Математическая статистика

Конспект лекций

Ростов-на-Дону 2021

§ 1. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

Оглавление.

Понятие случайной величины является основным в теории вероятностей и ее приложениях. Случайными величинами, например, является число выпавших очков при однократном бросании игральной кости, число распавшихся атомов радия за данный промежуток времени, число вызовов на телефонной станции за некоторый промежуток времени, отклонение от номинала некоторого размера детали при правильно налаженном технологическом процессе и т. д.

Таким образом, ***случайной величиной*** называется переменная величина, которая в результате опыта может принимать то или иное числовое значение.

В дальнейшем мы рассмотрим два типа случайных величин — ***дискретные*** и ***непрерывные.***

1. Дискретные случайные величины.

Рассмотрим случайную величину (случайные величины будем обозначать прописными буквами латинского алфавита: $X, Y, Z, \ldots$) X, возможные значения которой образуют конечную или бесконечную последовательность чисел $x_1, x_2, \ldots, x_n$. Такая случайная величина X называется ***дискретной (прерывной).***

На первый взгляд может показаться, что для задания дискретной случайной величины достаточно перечислить все ее возможные значения. В действительности это не так: случайные величины могут иметь одинаковые перечни возможных значений, а вероятности их – различные. Поэтому

для задания дискретной случайной величины недостаточно перечислить все возможные ее значения, нужно еще указать их вероятности.

Важнейшей характеристикой случайной величины служит ее распределение вероятностей.

Законом распределения дискретной случайной величины называют соответствие между возможными значениями и их вероятностями; его можно задать таблично, аналитически (в виде формулы) или графически.

При табличном задании закона распределения дискретной случайной величины первая строка таблицы содержит возможные значения, а вторая - их вероятности:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Приняв во внимание, что в одном испытании случайная величина принимает одно и только одно возможное значение, заключаем, что события $X = x_1$, $X = x_2$, …, $X = x_n$ образуют полную группу; следовательно, сумма вероятностей этих событий, т. е. сумма вероятностей второй строки таблицы, равна единице:

$$p_1 + p_2 + \ldots + p_n = 1.$$

$$P(X = x_k) = p_k, \qquad k = 1, 2, \ldots n, \qquad \sum_{k=1}^{n} p_k = 1 \tag{3.1}$$

Эту таблицу называют ***рядом распределения*** случайной величины X. Наглядно функцию ***p(x)*** можно изобразить в виде графика. Для этого возьмем прямоугольную систему координат на плоскости. По горизонтальной оси будем откладывать возможные значения случайной величины X, а по вертикальной оси - значения функции $p_i = P(x = x_i)$. График функции ***p(x)*** изображен на рис. 3.1. Если соединить точки этого графика прямолинейными отрезками, то

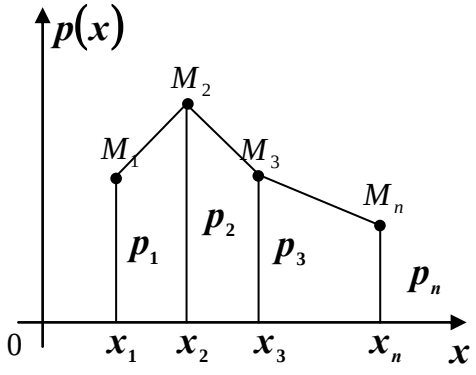

Рис. 3.1

получится фигура, которая называется ***многоугольником распределения***.

Пример 3.1. Пусть событие ***A*** — появление одного какого-либо очка при бросании игральной кости. Как мы знаем, вероятность выпадения какого-либо очка для всех цифр (1, 2, 3, 4, 5, 6) одинакова и равна ***P(A)=1/6***. Рассмотрим случайную величину ***X*** — число наступлений события *A* (т.е. число *m*) при десяти бросаниях игральной кости (т.е. ***n=10***). Значения функции ***p(x)*** (***закона распределения***) приведены в следующей таблице:

Значения X	0	1	2	3	4	5	6	…	10
Вероятности $p(x_i)$	0,162	0,323	0,291	0,155	0,054	0,013	0,002	…	0

$X=0$ означает, что цифра ***1*** (или любая другая из шести, могущих выпасть) при десяти бросаний кости не выпала ни разу. $X=1$ - цифра ***1*** при десяти бросаний выпала один раз. $X=2$ - два раза и т.д.

Вероятности $p(x_i)$ приведенные в таблице, вычислены по формуле Бернулли ($P_n(m)=\frac{n!}{m!\cdot(n-m)!}\cdot p^m\cdot q^{n-m}$) при ***n=10***. Для ***x>6*** они практически равны нулю:

$$P_{10}(0)=\frac{10!}{0!\cdot(10-0)!}\cdot\left(\frac{1}{6}\right)^0\cdot\left(\frac{5}{6}\right)^{10-0}=\frac{10!}{10!}\cdot 1\cdot\left(\frac{5}{6}\right)^{10}=0{,}162$$

$$P_{10}(1)=\frac{10!}{1!\cdot(10-1)!}\cdot\left(\frac{1}{6}\right)^1\cdot\left(\frac{5}{6}\right)^{10-1}=\frac{10!}{9!}\cdot\frac{1}{6}\cdot\left(\frac{5}{6}\right)^9=10\cdot\frac{1}{6}\cdot 0{,}19=0{,}32$$

$$P_{10}(2)=\frac{10!}{2!\cdot(10-2)!}\cdot\left(\frac{1}{6}\right)^2\cdot\left(\frac{5}{6}\right)^8=\frac{10!}{2\cdot 8!}\cdot\frac{1}{36}\cdot 0{,}232=0{,}291$$

$$P_{10}(3)=\frac{10!}{3!\cdot(10-3)!}\cdot\left(\frac{1}{6}\right)^3\cdot\left(\frac{5}{6}\right)^7=\frac{10!}{2\cdot 3\cdot 7!}\cdot\frac{1}{6^3}\cdot 0{,}232=0{,}155$$

$$P_{10}(4)=\frac{10!}{4!\cdot(10-4)!}\cdot\left(\frac{1}{6}\right)^4\cdot\left(\frac{5}{6}\right)^6=\frac{10!}{2\cdot3\cdot4\cdot6!}\cdot\frac{1}{6^4}\cdot0{,}3348=0{,}054$$

$$P_{10}(5)=\frac{10!}{5!\cdot(10-5)!}\cdot\left(\frac{1}{6}\right)^5\cdot\left(\frac{5}{6}\right)^5=\frac{10!}{5!\cdot5!}\cdot\frac{1}{6^5}\cdot0{,}4018=0{,}013$$

График функции p(x) изображен на рис. 3.2.

Это так называемый биноминальный закон распределения

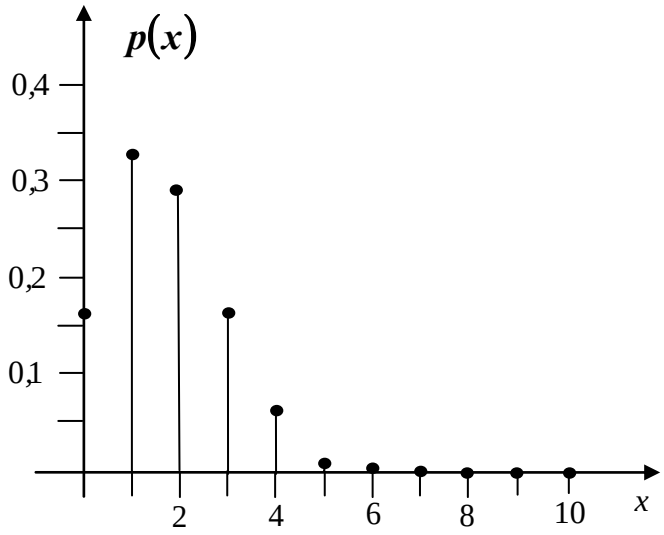

Рис. 3.2

2. Часто встречающиеся распределения дискретной случайной величины.

1. ***Закон распределения Бернулли***. Пусть случайная величина X это число, характеризующее наступления события A при одном испытании. При этом множество возможных значений X состоит из **2-х** чисел **0** и **1**: $x=0$, если событие A не произошло, и $x=1$ если событие A произошло. Таким образом: $p(0)=P(x=0)=P(\overline{A})=1-p=q \qquad p(1)=P(x=1)=P(A)=p$

x_i	0	1
p_i	q	p

Распределение Бернулли играет фундаментальную роль в теории вероятностей и математической статистики, являясь математической моделью опыта с двумя исходами.

Пусть, например, имеется партия некоторой продукции, в которой продукция без дефектов встречается с вероятностью p, а некачественная продукция с вероятностью $q = 1 - p$. Пусть случайная величина $X = 1$, если при выборе попалась качественная продукция и $X = 0$, если некачественная. Тогда случайная величина X будет иметь распределение Бернулли.

Пример 3.2. Случайная величина X — число очков, выпадающих при однократном бросании игральной кости. Возможные значения X — числа **1, 2, 3, 4, 5** и **6.** При этом вероятность того, что X примет любое из этих значений, одна и та же и равна **1/6**. Какой будет закон распределения?

Решение: Здесь закон распределения вероятностей есть функция ***p(x)=1/6*** для любого значения ***x*** из множества {**1, 2, 3, 4, 5, 6**}.

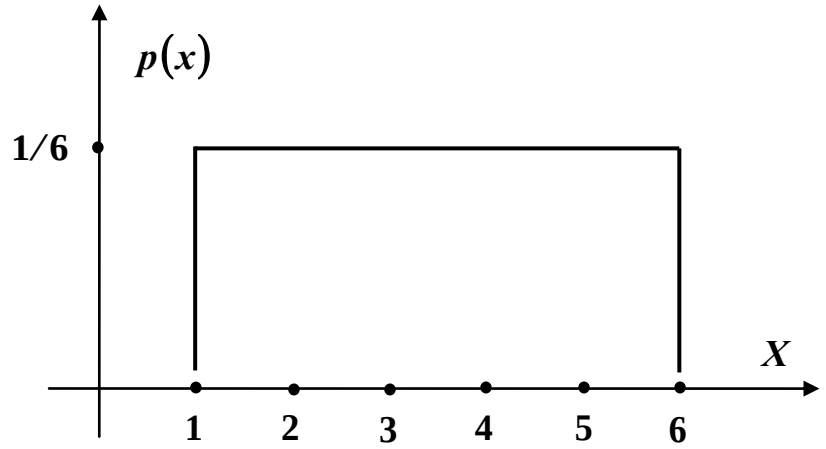

Рис. 3.3

График этого закона имеет вид, изображенный на рис. 3.3.

2. ***Биноминальный закон распределения***. Случайная величина X принимает значения: $0, 1, 2, 3, 4, 5, \ldots, n$, с вероятностью, определяемой по формуле Бернулли: $P(X = m) = P_n(m) = \dfrac{n!}{m!\,(n-m)!} p^m \cdot q^{n-m}$. Здесь p - постоянная вероятность того, что случайная величина X в серии n испытаний появится m раз.

x_i	0	1	2	…	k	…	n
p_i	$C_n^0 \cdot p^0 \cdot q^n$	$C_n^0 \cdot p^0 \cdot q^n$	$C_n^0 \cdot p^0 \cdot q^n$	…	$C_n^0 \cdot p^0 \cdot q^n$	…	$C_n^0 \cdot p^0 \cdot q^n$

Пример этого закона мы рассмотрели выше. График этого закона имеет вид, изображенный на рис. 3.4.

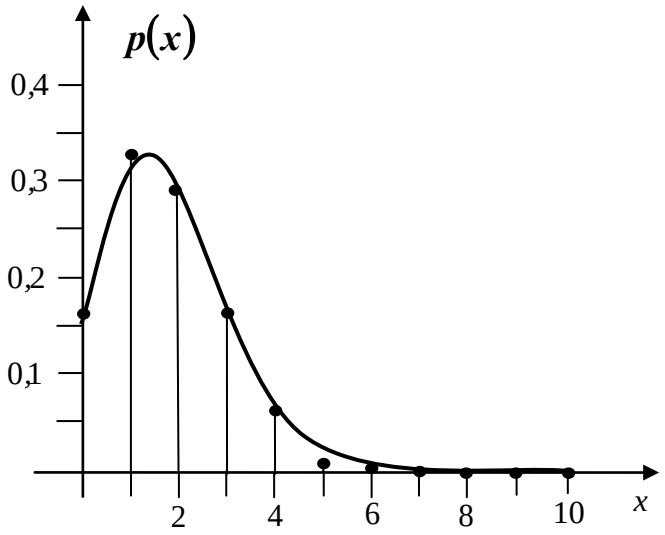

Рис. 3.4

3. ***Закон распределения Пуассона***. Случайная величина X принимает бесконечное счетное число значений: $0, 1, 2, 3, 4, 5, \ldots, k, \ldots,$ с вероятностью, определяющейся по формуле Пуассона:

$$p(k) = P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!} \qquad (3.2)$$

где $\lambda > 0$ — некоторая положительная постоянная - параметр распределения Пуассона..

В этом случае говорят, что случайная величина X распределена по ***закону Пуассона***. Заметим, что при $k = 0$ следует положить $0! = 1$.

Формулу Пуассона можно получить как предельный случай формулы Бернулли при неограниченном увеличении числа испытаний n и при стремлении к нулю вероятности $p = \frac{\lambda}{n}$.

Пример 18. На завод прибыла партия деталей в количестве **1000 шт.** Вероятность того, что деталь окажется бракованной, равна **0,001**. Какова

вероятность того, что среди прибывших деталей будет **5** бракованных?

Решение: Здесь $\lambda = np = 1000 \cdot 0{,}001 = 1$. По формуле (17) находим

$$P_{1000}(5) \approx \frac{1^5 \cdot e^{-1}}{5!} \approx 0{,}003$$

Распределение Пуассона часто встречается и в других задачах. Так, например, если телефонистка в среднем за один час получает N вызовов, то, как можно показать, вероятность $P(k)$ того, что в течение одной минуты она получит k вызовов, выражается формулой Пуассона, если положить $\lambda = \frac{N}{60}$.

$$P(k) = \frac{1}{k!} \cdot \left(\frac{N}{60}\right)^k \cdot e^{\left(-\frac{N}{60}\right)}$$

3. Функция распределения вероятностей случайной величины и ее свойства.

Рассмотрим некоторую функцию $F(x)$, определенную на всей числовой оси следующим образом: для каждого x значение этой функции $F(x)$ равно вероятности того, что дискретная случайная величина X примет значение, меньшее x, т. е.

$$F(x) = P(X < x) \qquad (3.3)$$

В этом случае эта функция называется ***функцией распределения вероятностей***, или кратко, ***функцией распределения***.

Пример 19. Найти функцию распределения случайной величины X, приведенной в **примере 17** (Случайная величина X — число очков, выпадающих при однократном бросании игральной кости).

Решение: Ясно, что если $x \le 1$, то $F(x)=0$, так как X не принимает значений, меньших единицы.

Если $1 < x \le 2$, то $F(x) = P(X < x) = P(X = 1) = \frac{1}{6}$.

Если $2 < x \le 3$, то $F(x) = P(X < x) = P(X < 3)$. Но событие $X < 3$ в данном случае является суммой двух несовместных событий: $X = 1$ и $X = 2$. Следовательно,

$$P(X < 3) = P(X = 1) + P(X = 2) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Итак, для $2 < x \le 3$ имеем ***F(x)=1/3.*** Аналогично вычисляются значения функции в промежутках $3 < x \le 4$, $4 < x \le 5$ и $5 < x \le 6$. Наконец, если $x > 6$ то ***F(x)=1***, так как в этом случае любое возможное значение X ***(1, 2, 3, 4, 5, 6)*** меньше, чем ***x.*** График функции ***F(x)*** изображен на рис. 3.5.

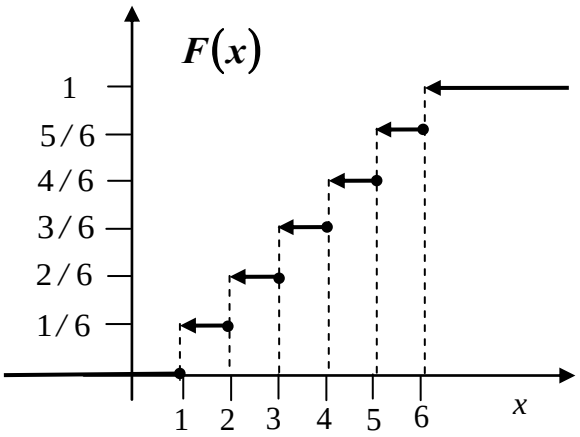

Рис. 3.5.

Зная функцию распределения ***F(x),*** легко найти вероятность того, что случайная величина X удовлетворяет неравенствам $x_1 \le X < x_2$.

Действительно, рассмотрим событие, заключающееся в том, что случайная величина примет значение меньшее x_2. Это событие распадается на сумму двух несовместных событий: 1) случайная величина X принимает значения, меньшие x_1, т.е. $X < x_1$; 2) случайная величина X принимает значения, удовлетворяющие неравенствам $x_1 \le X < x_2$. Используя аксиому сложения, получаем:

$$P(X < x_2) = P(X < x_1) + P(x_1 \le X < x_2).$$

Отсюда $P(x_1 \le X < x_2) = P(X < x_2) - P(X < x_1)$. Но по определению функции распределения ***F(x)*** [см. формулу (3.3)], имеем $P(X < x_2) = F(x_2)$, $P(X < x_1) = F(x_1)$; следовательно,

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1) \qquad (3.4)$$

Таким образом, ***вероятность попадания дискретной случайной***

величины в интервал $x_1 \le X < x_2$ равна приращению функции распределения на этом интервале.

Рассмотрим основные свойства функции распределения.

1°. ***Функция распределения является неубывающей.***

В самом деле, пусть $x_1 < x_2$. Так как вероятность любого события неотрицательна, то $P(x_1 \le X < x_2)$. Поэтому из формулы (3.4) следует, что $F(x_2) - F(x_1) \ge 0$, т.е. $F(x_2) > F(x_1)$.

2°. ***Значения функции распределения удовлетворяют неравенствам*** $0 \le F(x) \le 1$.

Это свойство вытекает из того, что ***F(x)*** определяется как вероятность [см. формулу (3.3)]. Ясно, что

$$F(-\infty) = 0 \text{ и } F(+\infty) = 1.$$

Здесь и в дальнейшем введены обозначения: $F(-\infty) = \lim_{x \to -\infty} F(x)$, $F(+\infty) = \lim_{x \to +\infty} F(x)$.

3°. ***Вероятность того, что дискретная случайная величина X примет одно из возможных значений x_i, равна скачку функции распределения в точке x_i.***

Действительно, пусть x_i - значение, принимаемое дискретной случайной величиной, и $\Delta x > 0$. Полагая в формуле (3.4) $x_1 = x_i$, $x_2 = x_i + \Delta x$, получим

$$P(x_i \le X < x_i + \Delta x) = F(x_i + \Delta x) - F(x_i)$$

В пределе при $\Delta x \to 0$ вместо вероятности попадания случайной величины на интервал $x_i \le X < x_i + \Delta x$ получим вероятность того, что величина X примет данное значение x_i:

$$\lim_{\Delta x \to 0} P(x_i \le X < x_i + \Delta x) = P(X = x_i) = p(x_i)$$

С другой стороны, получаем $\lim_{\Delta x \to 0} P(x_i + \Delta x) = F(x_i + 0)$, т.е. предел функции ***F(x)*** справа, так как $\Delta x > 0$. Следовательно, в пределе формула примет вид

$$p(x_i) = F(x_i + 0) - F(x_i) = F(x_i + 0) - F(x_i - 0)$$

т.е. значение $p(x_i)$ равно скачку функции $F(x_i)$. Можно показать, что $F(x_i) = F(x_i - 0)$, т.е. что функция ***F(x)*** непрерывна слева в точке x_i. Это свойство наглядно иллюстрируется на рис.4.

4. Непрерывные случайные величины.

Кроме дискретных случайных величин, возможные значения которых образуют конечную или бесконечную последовательность чисел, не заполняющих сплошь никакого интервала, часто встречаются случайные величины, возможные значения которых образуют некоторый интервал.

Примером такой случайной величины может служить отклонение от номинала некоторого размера детали при правильно налаженном технологическом процессе.

Такого рода, случайные величины не могут быть заданы с помощью закона распределения вероятностей $p(x)$. Однако их можно задать с помощью функции распределения вероятностей ***F(x).*** Эта функция определяется точно так же, как и в случае дискретной случайной величины:

$$F(x) = P(X < x)$$

Таким образом, и здесь функция ***F(x)*** определена на всей числовой оси, и ее значение в точке x равно вероятности того, что случайная величина примет значение, меньшее, чем x.

Формула (3.4) и свойства 1° и 2° справедливы для функции распределения любой случайной величины. Доказательство проводится аналогично случаю дискретной величины.

Случайная величина X называется ***непрерывной***, если для нее существует неотрицательная кусочно-непрерывная функция (функция называется кусочно-непрерывной на всей числовой оси, если она на любом сегменте или непрерывна, или имеет конечное число точек разрыва I рода) $\varphi(x)$, удовлетворяющая для любых значений x равенству

$$F(x) = \int_{-\infty}^{x} \varphi(t)\,dt \qquad (3.5)$$

Функция $\varphi(t)$ называется ***плотностью распределения вероятностей***, или кратко, ***плотностью распределения***. Если $x_1 < x_2$, то на основании формул (3.3) и (3.4) имеем

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1) = \int_{-\infty}^{x_2} \varphi(t)\,dt - \int_{-\infty}^{x_1} \varphi(t)\,dt = \int_{x_1}^{x_2} \varphi(t)\,dt \qquad (3.6)$$

Исходя из геометрического смысла интеграла как площади, можно сказать, что вероятность выполнения неравенств $x_1 \le X \le x_2$ равна площади криволинейной трапеции с основанием $[x_1, x_2]$, ограниченной сверху кривой $y = \varphi(x)$ (рис. 3.6).

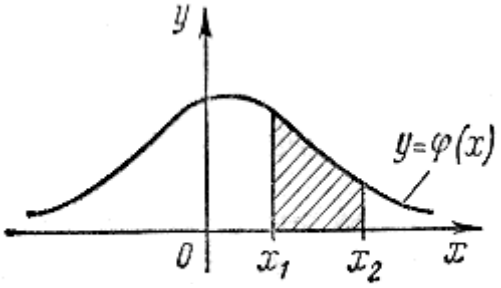

Рис. 3.6

Так как $F(+\infty) = P(X < +\infty) = 1$, а на основании формулы (26) $F(x) = \int_{-\infty}^{x} \varphi(t)\,dt$, то

$$\int_{-\infty}^{+\infty} \varphi(t)\,dt = 1 \qquad (3.7)$$

Далее, пользуясь формулой (3.5), найдем $F'(x)$, производную от $F(x)$, как производную от интеграла по переменной верхней границе, считая плотность распределения $\varphi(x)$ непрерывной. Правило дифференцирования интеграла с переменной верхней границей, выведенное в случае конечной нижней границы, остается справедливым и для интегралов с бесконечной нижней границей. В самом деле,

$$\frac{d}{dx}\left(\int_{-\infty}^{x} \varphi(t)dt\right) = \frac{d}{dx}\left(\int_{-\infty}^{a} \varphi(t)dt + \int_{a}^{x} \varphi(t)dt\right) = 0 + \varphi(x) = \varphi(x)$$

Так как интеграл $\int_{-\infty}^{a} \varphi(t)dt$ есть величина постоянная. Таким образом

$$F'(x) = \frac{d}{dx}\left(\int_{-\infty}^{x} \varphi(t)\,dt\right) = \varphi(x) \qquad (3.8)$$

Заметим, что для непрерывной случайной величины функция распределения ***F(x)*** непрерывна в любой точке ***x,*** где функция $\varphi(x)$ непрерывна. Это следует из того, что ***F(x)*** в этих точках дифференцируема.

Далее, на основании формулы (3.6), полагая $x_1 = x$, $x_2 = x + \Delta x$, имеем:

$$P(x \le X < x + \Delta x) = F(x + \Delta x) - F(x) = \Delta F(x)$$

В силу непрерывности функции ***F(x)*** получим, что $\lim\limits_{\Delta x \to 0} \Delta F(x) = 0$. Следовательно $\lim\limits_{\Delta x \to 0} P(x \le X < x + \Delta x) = P(X = x) = 0$. Таким образом,

вероятность того, что непрерывная случайная величина может принять любое отдельное значение x, равна нулю.

Отсюда следует, что события, заключающиеся в выполнении каждого из неравенств: $x_1 \le X < x_2$; $x_1 < X \le x_2$; $x_1 \le X \le x_2$; $x_1 < X < x_2$; имеют одинаковую вероятность, т.е.

$$P(x_1 \le X < x_2) = P(x_1 < X \le x_2) = P(x_1 \le X \le x_2) = P(x_1 < X < x_2)$$

В самом деле, например,

$$P(x_1 \le X < x_2) = P(X = x_1) + P(x_1 < X < x_2) = P(x_1 < X < x_2)$$

так как $P(X = x_1) = 0$.

Замечание.

Как мы знаем, если событие невозможно, то вероятность его наступления равна нулю. При классическом определении вероятности, когда число исходов испытаний конечно, имеет место и обратное предложение: если вероятность события равна нулю, то событие невозможно, так как в этом случае ему не благоприятствует ни один из исходов испытания.

В случае непрерывной случайной величины число возможных ее значений бесконечно. Вероятность того, что эта величина примет какое-либо конкретное значение x_i как мы видели, равна нулю. Однако отсюда не следует, что это событие невозможно, так как в результате испытания случайная величина может, в частности, принять значение x_i.

Поэтому в случае непрерывной случайной величины имеет смысл говорить о **вероятности попадания случайной величины в интервал**, а не о вероятности того, что она примет какое-то конкретное значение.

Так, например, при изготовлении валика нас не интересует вероятность того, что его диаметр будет равен номиналу. Для нас важна вероятность того, что диаметр валика не выходит из поля допуска.

Пример 20. Плотность распределения непрерывной случайной величины задана следующим образом:

$$\varphi(x)=\begin{cases} 0, & \text{если} \quad x<0 \\ \dfrac{3}{32}\cdot(4x-x^2), & \text{если} \quad 0\le x\le 4 \\ 0, & \text{если} \quad x>4 \end{cases}$$

График функции $\varphi(x)$ представлен па рис. 3.7.

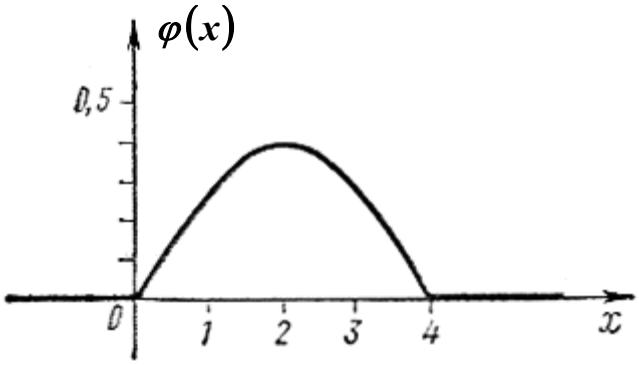

Рис. 3.7

Определить: 1) вероятность того, что случайная величина X примет значение, удовлетворяющее неравенствам $-2\le X\le 3$. 2) Найти функцию распределения заданной случайной величины.

Решение: Используя формулу (3.6), имеем:

$$P(-2\le X\le 3)=\int_{-2}^{3}\varphi(x)dx=\int_{-2}^{0}\varphi(x)dx+\int_{0}^{3}\varphi(x)dx=\int_{-2}^{0}0\,dx+\int_{0}^{3}\frac{3}{32}(4x-x^2)\,dx=\frac{27}{32}$$

По формуле (3.6) находим функцию распределения $F(x)$ для заданной случайной величины. Если $-\infty<x\le 0$, то $F(x)=\int_{-\infty}^{x}\varphi(t)dt=\int_{-\infty}^{0}0\cdot dt=0$.

Если $0<x\le 4$, то

$$F(x)=\int_{-\infty}^{x}\varphi(t)dt=\int_{-\infty}^{0}\varphi(t)\cdot dt+\int_{0}^{x}\varphi(t)\cdot dt=\int_{-\infty}^{0}0\cdot dt+\int_{0}^{x}\frac{3}{32}(4t-t^2)\cdot dt=\frac{6x^2-x^3}{32}$$

Если $x>4$, то

$$F(x)=\int_{-\infty}^{x}\varphi(t)dt=\int_{-\infty}^{0}\varphi(t)\cdot dt+\int_{0}^{4}\varphi(t)\cdot dt+\int_{4}^{x}\varphi(t)\cdot dt=$$

$$=\int_{-\infty}^{0}0\cdot dt+\int_{0}^{4}\frac{3}{32}(4t-t^2)\cdot dt+\int_{4}^{x}0\cdot dt=1$$

Итак,

$$F(x)=\begin{cases} 0, & \text{если} \quad x<0 \\ \dfrac{6x^2-x^3}{32}, & \text{если} \quad 0\le x\le 4 \\ 1, & \text{если} \quad x>4 \end{cases}$$

График функции *F(x)* изображен на рис. 3.8.

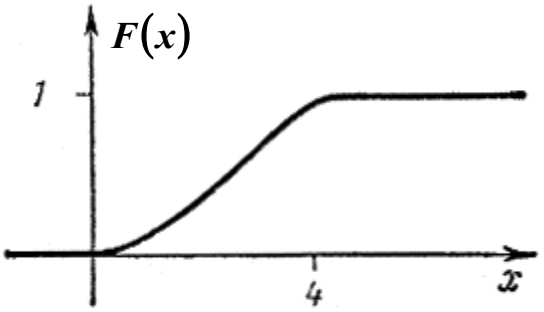

Рис. 3.8.

Следующие три пункта посвящены часто встречающимся на практике распределениям непрерывных случайных величин — равномерному, экспоненциальному и нормальному распределениям.

5. Равномерное распределение.

Пусть сегмент $[a,b]$ на оси Ox есть шкала некоторого прибора. Допустим, что вероятность попадания указателя в некоторый отрезок шкалы пропорциональна длине этого отрезка и не зависит от места отрезка на шкале. Отметка указателя прибора есть случайная величина X могущая принять любое значение из сегмента $[a,b]$. Поэтому $P(a\le X\le b)=1$. Если, далее, x_1 и x_2 $(x_1<x_2)$ - две любые отметки на шкале, то согласно условию имеем $P(x_1\le X\le x_2)=k\cdot(x_2-x_1)$, где k - коэффициент пропорциональности, не зависящий от x_1 и x_2, а разность x_2-x_1, - длина сегмента $[x_1, x_2]$. Так как при $x_1=a$ и $x_2=b$ имеем $P(a\le X\le b)=1$, то $k\cdot(b-a)=1$, откуда $k=\dfrac{1}{b-a}$. Таким образом

$$P(x_1 \le X \le x_2) = \frac{x_2 - x_1}{b - a} \tag{3.9}$$

Теперь легко найти функцию $F(x)$ распределения вероятностей случайной величины X. Если $x \le a$, то $F(x) = P(X < x) = 0$ так как X не принимает значений, меньших a. Пусть теперь $a < x \le b$. По аксиоме сложения вероятностей $P(X < x) = P(X < a) + P(a \le X < x)$. Согласно формуле (3.9), в которой принимаем $x_1 = a$ и $x_2 = x$, имеем $P(a \le X \le x) = \frac{x - a}{b - a}$. Так как $P(X < a) = 0$, то при $a < x \le b$ получаем $F(x) = P(X < x) = \frac{x - a}{b - a}$. Наконец, если $x > b$, то $F(x) = 1$, так как значения X лежат на сегменте $[a, b]$ и, следовательно, не превосходят b. Итак, приходим к следующей функции распределения:

$$F(x) = \begin{cases} 0, & \text{если} \quad x \le a \\ \dfrac{x - a}{b - a}, & \text{если} \quad a < x \le b \\ 1, & \text{если} \quad x > b \end{cases}.$$

График функции $F(x)$ представлен на рис. 3.9.

Плотность распределения вероятностей найдем по формуле (3.8). Если $x < a$ или $x > b$, то $\varphi(x) = F'(x) = 0$. Если $a < x < b$, то

$$\varphi(x) = F'(x) = \left(\frac{x - a}{b - a}\right)' = \frac{1}{b - a}.$$

Таким образом,

$$\varphi(x) = \begin{cases} 0, & \text{если} \quad x < a \\ \dfrac{1}{b - a}, & \text{если} \quad a < x < b \\ 1, & \text{если} \quad x > b \end{cases} \tag{3.10}$$

График функции $\varphi(x)$ изображен на рис. 3.10. Заметим, что в точках a и b функция $\varphi(x)$ терпит разрыв.

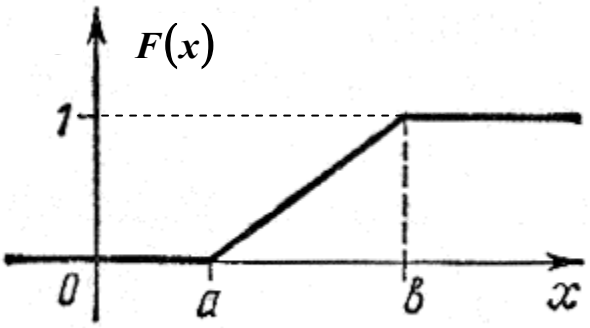

Рис. 3.9

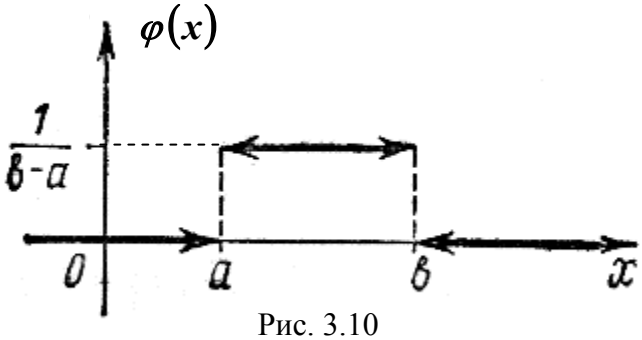

Рис. 3.10

Величина, плотность распределения которой задана формулой (3.10), называется ***равномерно распределенной*** случайной величиной.

6. Нормальное распределение.

Говорят, что случайная величина ***X нормально распределена*** или подчиняется ***закону распределения Гаусса***, если ее плотность распределения $\varphi(x)$ имеет вид

$$\varphi(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-a)^2/(2\cdot\sigma^2)} \qquad (3.11)$$

где a - любое действительное число, а $\sigma>0$. Смысл параметров a и σ будет установлен в дальнейшем. Исходя из связи между плотностью распределения $\varphi(x)$ и функцией распределения ***F(x)*** [см. формулы (3.5, 3.8)], имеем

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-(t-a)^2/(2\cdot\sigma^2)}dt$$

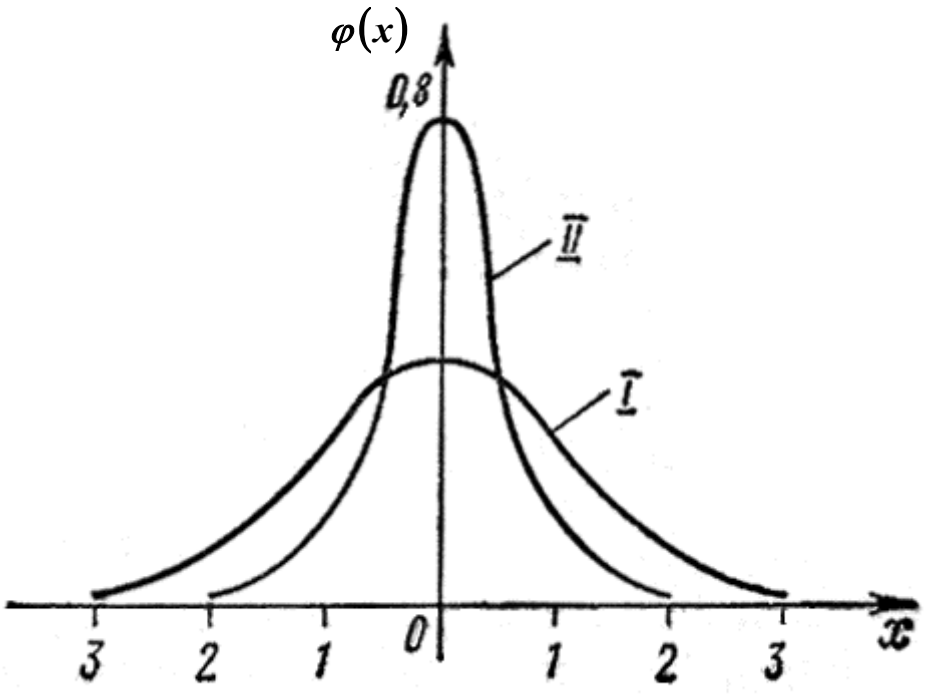

Рис. 3.11.

График функции $\varphi(x)$ симметричен относительно прямой $x = a$. Несложные исследования показывают, что функция $\varphi(x)$ достигает максимума при $x = a$, а ее график имеет точки перегиба при $x_1 = a + \sigma$ и $x_2 = a - \sigma$. При $x \to \pm\infty$ график функции асимптотически приближается к оси Ox. Можно показать, что при увеличении σ кривая плотности распределения становится более пологой. Наоборот, при уменьшении σ график плотности распределения сжимается к оси симметрии. При $a = 0$ осью симметрии является ось Oy. На рис. 3.11 изображены два графика функции $y = \varphi(x)$. График I соответствует значениям $a = 0$, $\sigma = 1$, а график II - значениям $a = 0$, $\sigma = 1/2$.

Покажем, что функция $\varphi(x)$ удовлетворяет условию (3.7), т.е. при любых a и σ выполняется соотношение

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/(2\cdot\sigma^2)} dx = 1$$

В самом деле, сделаем в этом интеграле замену переменной, полагая $\frac{x-a}{\sigma} = t$. Тогда

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/(2\sigma^2)} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-t^2/2} dt$$

В силу четности подынтегральной функции имеем

$$\int_{-\infty}^{+\infty} e^{-t^2/2} dt = 2\int_{0}^{+\infty} e^{-t^2/2} dt .$$

Следовательно,

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/(2\sigma^2)} dx = \frac{2}{\sqrt{2\pi}} \int_{0}^{+\infty} e^{-t^2/2} dt$$

Но, $$\int_{0}^{+\infty} e^{-t^2/2} dt = \sqrt{\frac{\pi}{2}} .$$

В результате получим

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(x-a)^2/(2\sigma^2)} dx = 1 \qquad (3.12)$$

Найдем вероятность $P(x_1 < X < x_2)$. По формуле (3.6) имеем

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} \varphi(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-(x-a)^2/(2\sigma^2)} dx$$

Сделаем в этом интеграле замену переменной, снова полагая $\frac{x-a}{\sigma} = t$.

Тогда $x = a + \sigma \cdot t, \quad dx = \sigma \cdot dt$, и

$$P(x_1 < X < x_2) = \frac{1}{\sqrt{2\pi}} \int_{(x_1-a)/\sigma}^{(x_2-a)/\sigma} e^{-t^2/2} dt \qquad (3.13)$$

Как мы знаем, интеграл $\int e^{-t^2/2} dt$ не берется в элементарных функциях. Поэтому для вычисления определенного интеграла (3.13) вводится функция, которую мы определяли раньше [формула (2.9)] :

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-t^2/2} dt \qquad (3.14)$$

называемая ***интегралом вероятностей***.

Для этой функции составлены таблицы ее значений для различных значений аргумента (см. табл. II Приложения). Используя формулу (3.13) получим:

$$P(x_1 < X < x_2) = \frac{1}{\sqrt{2\pi}} \int_{(x_1-a)/\sigma}^{(x_2-a)/\sigma} e^{-t^2/2}dt = \frac{1}{\sqrt{2\pi}} \int_{(x_1-a)/\sigma}^{0} e^{-t^2/2}dt + \frac{1}{\sqrt{2\pi}} \int_{0}^{(x_2-a)/\sigma} e^{-t^2/2}dt =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{(x_2-a)/\sigma} e^{-t^2/2}dt - \frac{1}{\sqrt{2\pi}} \int_{0}^{(x_1-a)/\sigma} e^{-t^2/2}dt = \Phi\left(\frac{x_2-a}{\sigma}\right) - \Phi\left(\frac{x_1-a}{\sigma}\right)$$

Итак,

$$P(x_1 < X < x_2) = \Phi\left(\frac{x_2-a}{\sigma}\right) - \Phi\left(\frac{x_1-a}{\sigma}\right) \qquad (3.15)$$

Легко показать, что функция $\Phi(x)$ (*интеграл вероятностей*) обладает следующими свойствами.

1°. $\Phi(0)=0$

2°. $\Phi(\infty) = \frac{1}{\sqrt{2\pi}} \int_{0}^{+\infty} e^{-t^2/2}dt = \frac{1}{2}$;

при $|x| \geq 4$ величина $|\Phi(x)|$ практически равна $1/2$ (см. табл. II).

3°. $\Phi(-x) = -\Phi(x)$,

т.е. интеграл вероятностей является нечетной функцией.

График функции $\Phi(x)$ изображен на рис. 3.12.

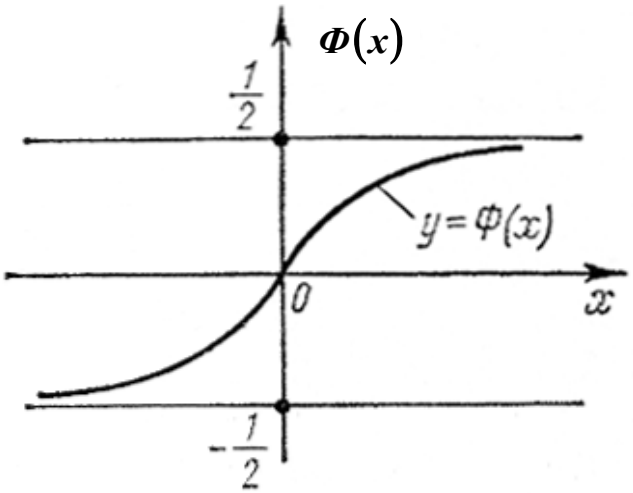

Рис. 3.12

Таким образом, если случайная величина X нормально распределена с параметрами a и σ, то вероятность того, что случайная величина удовлетворяет неравенствам $x_1 < X < x_2$, определяется соотношением (3.15).

Пусть $X > 0$. Найдем вероятность того, что нормально распределенная случайная величина X отклонится от параметра a по абсолютной величине не более чем на ε, т.е. рассмотрим неравенство - $P(|X - a|) < \varepsilon$.

Так как неравенство $|X - a| < \varepsilon$ равносильно неравенствам $a - \varepsilon < X < a + \varepsilon$, то полагая в соотношении (3.15) $x_1 = a - \varepsilon$, $x_2 = a + \varepsilon$ получим

$$P(|X - a| < \varepsilon) = P(a - \varepsilon < X < a + \varepsilon) = \Phi\left(\frac{a + \varepsilon - a}{\sigma}\right) - \Phi\left(\frac{a - \varepsilon - a}{\sigma}\right) = \Phi\left(\frac{\varepsilon}{\sigma}\right) - \Phi\left(-\frac{\varepsilon}{\sigma}\right)$$

Вследствие того, что интеграл вероятностей - нечетная функция, имеем

$$P(|X - a| < \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sigma}\right) \qquad (3.16)$$

Пример 1. Пусть случайная величина X подчиняется нормальному закону распределения вероятностей с параметрами $a = 0$, $\sigma = 2$.

Определить:

1) $P(-2 < X < 3)$; 2) $P(|X| < 0{,}1)$;

Решение: 1) Используя формулу (3.15), имеем

$$P(-2 < X < 3) = \Phi\left(\frac{3 - 0}{2}\right) - \Phi\left(\frac{-2 - 0}{2}\right) = \Phi(1{,}5) - \Phi(-1) = \Phi(1{,}5) + \Phi(1)$$

Из таблицы II находим, что $\Phi(1) = 0{,}34134$, $\Phi(1{,}5) = 0{,}43319$.

Следовательно

$$P(-2 < X < 3) = 0{,}43319 + 0{,}34134 = 0{,}77453$$

2) Так как $a = 0$, то $|\xi| = |\xi - a|$. По формуле (3.16) находим

$$P(|X| < 0{,}1) = 2\Phi\left(\frac{0{,}1}{2}\right) = 2\Phi(0{,}05) = 2 \cdot 0{,}01994 = 0{,}03988$$

Пример 2. В каких пределах должна изменяться случайная величина, подчиняющаяся нормальному закону распределения, чтобы $P(|X-a|<\varepsilon)=0{,}9973$.

Решение: По формуле (37) имеем

$$P(|X-a|<\varepsilon)=2\Phi\left(\frac{\varepsilon}{\sigma}\right)=0{,}9973.$$

Следовательно, $\Phi(\varepsilon/\sigma)=0{,}49865$. Из табл. II находим, что этому значению $\Phi(\varepsilon/\sigma)$ соответствует $\varepsilon/\sigma=3$, откуда $\varepsilon=3\sigma$.

Из последнего примера следует, что если случайная величина подчиняется нормальному закону распределения, то можно утверждать с вероятностью, равной ***0,9973***, что случайная величина находится в интервале $[a-3\sigma, a+3\sigma]$. Так как данная вероятность близка к единице, то можно считать, что значения нормально распределенной случайной величины практически не выходят за границы интервала $[a-3\sigma, a+3\sigma]$. Этот факт называют ***правилом трех сигм.***

Аналогично можно посчитать, что вероятность того, что случайная величина, распределенная по нормальному закону, заключена в интервале $[a-2\sigma, a+2\sigma]$, равна ***95,44 %***. Соответственно в интервале $[a-\sigma, a+\sigma]$ равна ***67,26 %***. То есть:

$$P(|X-a|<3\sigma)=2\Phi(3)=0{,}9973$$

$$P(|X-a|<2\sigma)=2\Phi(2)=0{,}9544$$

$$P(|X-a|<\sigma)=2\Phi(1)=0{,}6826$$

Данные условия наглядно изображены на рис. 3.13.

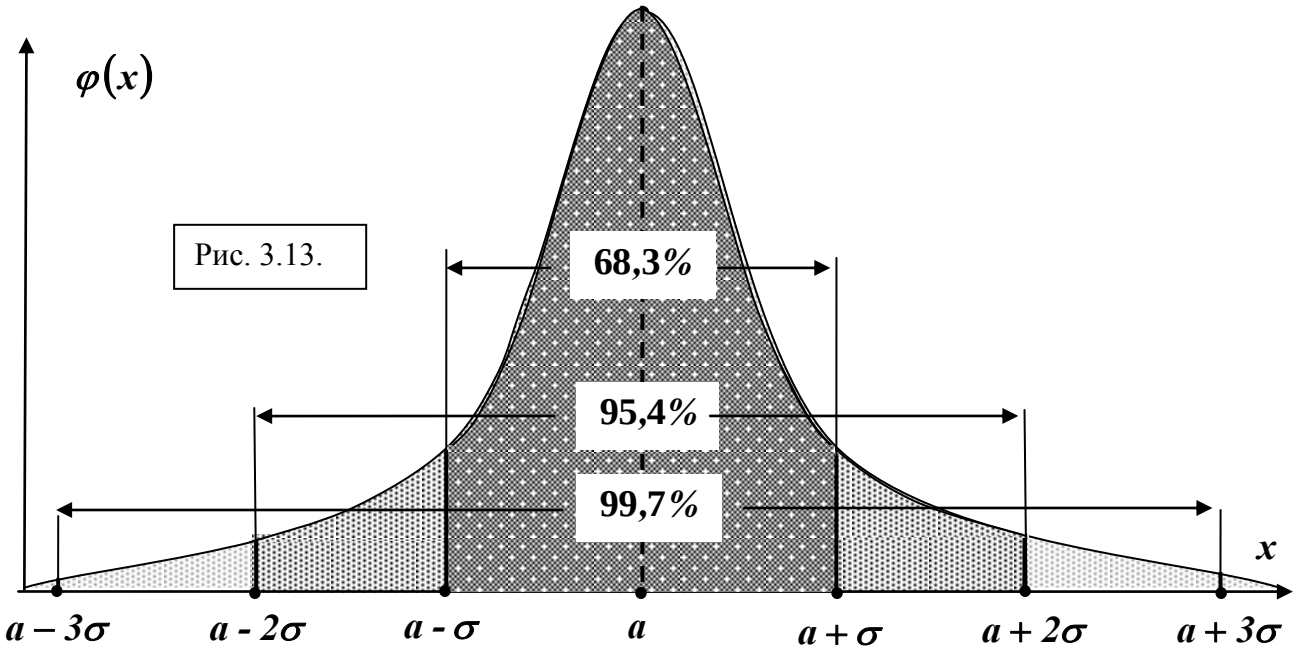

Рис. 3.13.

7. Экспоненциальное распределение.

Непрерывная случайная величина имеет экспоненциальное (показательное) распределение с параметром $\lambda > 0$, если

$$\varphi(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

Функция распределения этой случайной величины будет равна:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt = \int_{0}^{x} \lambda \cdot e^{-\lambda \cdot t}\, dt = \{\varphi(t) = 0, \quad если \quad t \le 0\} =$$

$$= -\int_{0}^{x} e^{-\lambda \cdot t}\, d(-\lambda t) = -e^{-\lambda \cdot t}\Big|_{0}^{x} = -e^{-\lambda \cdot x} + 1 = 1 - e^{-\lambda \cdot x}$$

Соответствующие графики изображены на рис. 3.14.

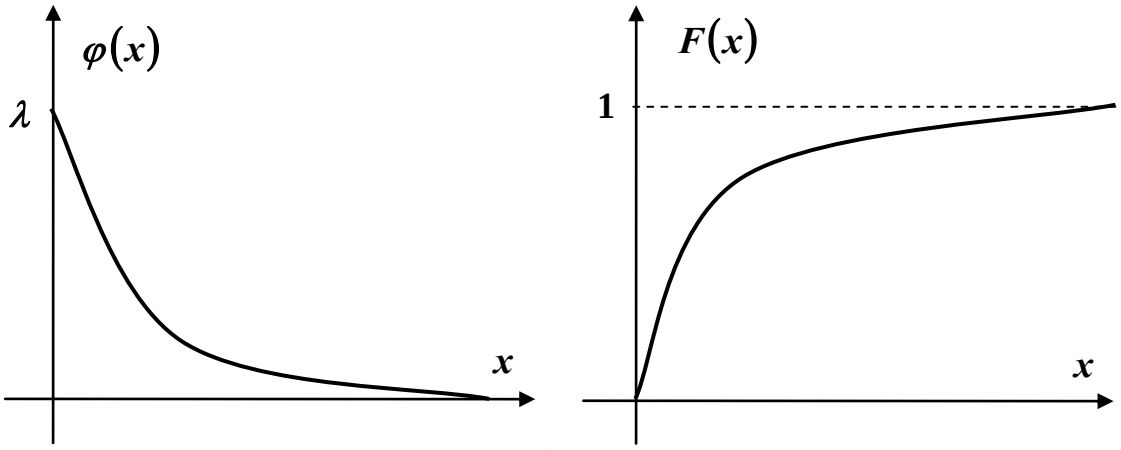

Рис. 3.14.

Продолжительность безотказной работы многих технических устройств, а также время задержки вылетов самолета, по вине технических служб аэропорта удовлетворительно описываются соответствующими экспоненциальными распределениями.

8. Двумерные случайные величины.

Часто приходится решать задачи, в которых рассматриваются события, описываемые не одной, а несколькими — в частности, двумя случайными величинами. Так если станок-автомат штампует цилиндрические валики, то диаметр валика X_1 и его высота X_2, образуют систему двух случайных величин (X_1, X_2).

Двумерной случайной величиной называют систему из двух случайных величин (X_1, X_2), для которой определена вероятность $P[(X_1 < x), (X_2 < y)]$ совместного выполнения неравенств $X_1 < x$ и $X_2 < y$, где x и y - любые действительные числа.

Функция двух переменных

$$F(x, y) = P[(X_1 < x), (X_2 < y)] \qquad (3.17)$$

определенная для любых x и y, называется ***функцией распределения*** системы двух случайных величин (X_1, X_2).

Двумерная случайная величина (X_1, X_2) называется ***дискретной***, если X_1 и X_2 - дискретные величины.

Будем рассматривать X_1 и X_2 как декартовы координаты точки на плоскости. Точка $M(X_1, X_2)$ может занимать то или иное положение на плоскости OX_1X_2. Тогда функция распределения даст вероятность того, что случайная точка $M(X_1, X_2)$ попадает в область σ, изображенную на рис. 3.15.

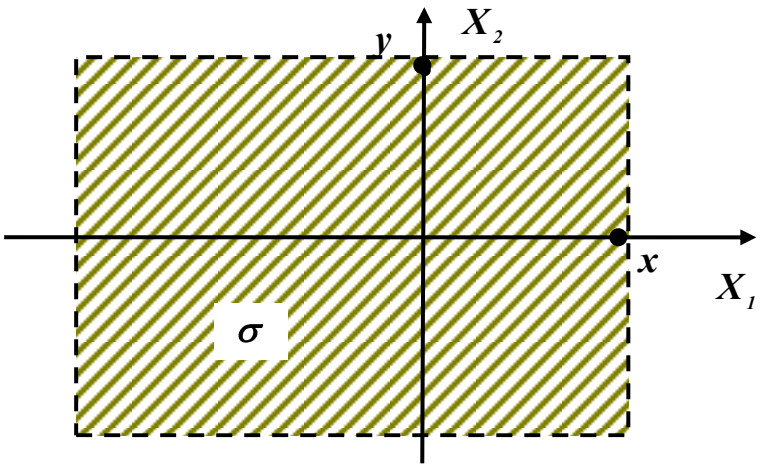

Рис. 3.15.

Пусть возможные значения X_1 и X_2 образуют, например, конечные последовательности $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_S$. Возможные значения двумерной случайной величины (X_1, X_2) имеют вид (x_i, y_j), где $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, s$. Обозначим через p_{ij} вероятность того, что $(X_1, X_2) = (x_i, y_j)$:

$$p_{ij} = P[(X_1 = x_i), (X_2 = y_j)]$$

Функция распределения ***F(x, y)*** имеет вид

$$F(x, y) = \sum_i \sum_j p_{ij}$$

где двойная сумма распространена на те i и j, для которых

$$x_i < x \quad \text{и} \quad y_j < y.$$

Двумерную случайную величину (X_1, X_2) так же, как и одномерную, можно задавать таблицей. Первая строка таблицы содержит возможные значения случайной величины X_1, а первый столбец — возможные значения X_2. В остальных клетках таблицы указаны соответствующие вероятности, причем их сумма всегда равна единице. В качестве примера рассмотрим двумерную случайную величину, заданную следующей таблицей:

$X_2 \backslash X_1$	-1	0	1
	$p_{11}=0,05$	$p_{12}=0,20$	$p_{13}=0,30$
	$p_{21}=0,10$	$p_{22}=0,20$	$p_{23}=0,15$

Сумма всех вероятностей

$$\sum_{i=1}^{3}\sum_{j=1}^{2} p_{ij} = p_{11}+p_{12}+p_{13}+p_{21}+p_{22}+p_{23} = 0,05+0,20+0,30+0,10+0,20+0,15 = 1,00$$

Две дискретные случайные величины X_1 и X_2 называются независимыми, если для всех пар i, j выполняется соотношение

$$p_{ij} = P\left[(X_1 = x_i),(X_2 = y_j)\right] = P(X_1 = x_i)\cdot P(X_2 = y_j)$$

Пример 1. Две игральные кости бросают по одному разу. Обозначим через X_1 число очков, выпавшее на первой кости, а через X_2 — на второй; тогда (X_1, X_2) — двумерная дискретная величина. Покажем, что величины X_1 и X_2 независимы.

Решение: Так как каждая из величин X_1 и X_2 независимо друг от друга может принимать 6 различных значений, то число различных значений двумерной случайной величины (X_1, X_2) равно 36. Все эти значения, очевидно, равновероятны. Поэтому

$$P\left[(X_1 = x_i),(X_2 = y_j)\right] = \frac{1}{36}$$

С другой стороны, $P(X_1 = x_i) = \frac{1}{6}$ и $P(X_2 = y_j) = \frac{1}{6}$. Таким образом:

$$P\left[(X_1 = x_i),(X_2 = y_j)\right] = P(X_1 = x_i)\cdot P(X_2 = y_j) = \frac{1}{36}$$

Так как вероятность появления равна произведению их вероятностей, то величины независимы.

Двумерная величина (X_1, X_2) называется ***непрерывной***, если существует такая непрерывная неотрицательная функция $\varphi(x, y)$ двух переменных, что вероятность того, что точка $M(X_1, X_2)$ содержится в некоторой области σ плоскости OX_1X_2, равна двойному интегралу от функции $\varphi(x, y)$ по области σ:

$$P\left(M\left(X_1, X_2\right) \in \sigma\right) = \iint_{\sigma} \varphi(x, y)\, dx\, dy \qquad (3.18)$$

Функция $\varphi(x, y)$ называется плотностью распределения вероятностей системы двух величин X_1 и X_2. Отсюда, в частности, следует, что если область σ имеет вид, изображенный на рис. 14, то функцию распределения системы случайных величин можно записать следующим образом:

$$F(x, y) = P\left[\left(X_1 < x\right) \cdot \left(X_2 < y\right)\right] = \iint_{\sigma} \varphi(x, y)\, dx\, dy = \int_{-\infty}^{x} dx \int_{-\infty}^{y} \varphi(x, y)\, dx\, dy \qquad (3.19)$$

Непрерывные случайные величины X_1 и X_2 называются ***независимыми***, если $\varphi(x, y) = \varphi_1(x) \cdot \varphi_2(y)$, где $\varphi_1(x)$ и $\varphi_2(y)$ - соответственно плотности распределения вероятностей случайных величин X_1 и X_2. В этом случае

$$F(x, y) = P\left[\left(X_1 < x\right), \left(X_2 < y\right)\right] = \int_{-\infty}^{x} dx \int_{-\infty}^{y} \varphi_1(x) \cdot \varphi_2(y)\, dx\, dy = \int_{-\infty}^{x} \varphi_1(x)\, dx \cdot \int_{-\infty}^{y} \varphi_2(y)\, dy$$

Зная функцию распределения ***F(x,y)*** двумерной случайной величины $\left(X_1, X_2\right)$, легко найти как функцию распределения, так и плотность распределения каждой из случайных величин X_1 и X_2 в отдельности.

Действительно, пусть ***F₁(x)*** - функция распределения случайной величины X_1. Тогда $F_1(x) = P\left(X < x\right)$. Так как в этом случае X_2 может принимать любое значение, то ясно, что

$$P\left(X_1 < x\right) = P\left[\left(X_1 < x\right), \left(-\infty < X_2 < +\infty\right)\right]$$

Следовательно, запишем:

$$F_1(x) = P\left(X_1 < x\right) = P\left[\left(X_1 < x\right), \left(-\infty < X_2 < +\infty\right)\right] = \int_{-\infty}^{x} dx \int_{-\infty}^{+\infty} \varphi(x, y)\, dy$$

Дифференцируя последнее равенство по ***x,*** согласно правилу дифференцирования интеграла по переменной верхней границе получим:

$$\varphi_1(x) = F_1'(x) = \int_{-\infty}^{+\infty} \varphi(x, y)\, dy \qquad (3.20)$$

Аналогичным образом получаем

$$F_2(x) = P\left(X_2 < y\right) = \int_{-\infty}^{y} dy \int_{-\infty}^{+\infty} \varphi(x, y)\, dx$$

и, следовательно,

$$\varphi_2(y)=\int_{-\infty}^{+\infty}\varphi(x,y)dx \qquad (3.21)$$

Таким образом, чтобы получить плотность распределения одной из составляющих двумерной случайной величины, надо проинтегрировать в границах от $-\infty$ до $+\infty$ плотность распределения системы $\varphi(x,y)$ по переменной, соответствующей другой случайной величине.

Пример 2. Двумерная случайная величина (X_1, X_2) имеет плотность распределения

$$\varphi(x,y)=\frac{1}{\pi^2\cdot(1+x^2)\cdot(1+y^2)}$$

Найти:

1) вероятность ***p*** попадания случайной точки $M(X_1,X_2)$ в квадрат изображенный на рис. 3.16;

2) функцию распределения ***F(x,y);***

3) плотности распределения каждой величины X_1 и X_2 в отдельности.

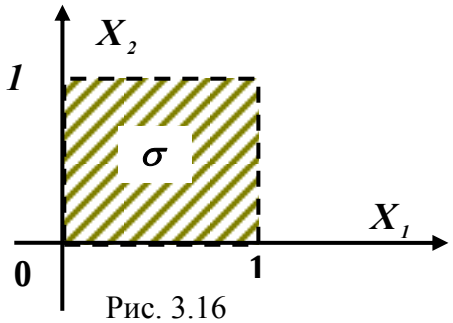

Рис. 3.16

Решение:

1) Вероятность ***p*** попадания случайной точки $M(X_1,X_2)$ в квадрат изображенный на рис. 3.16, согласно формуле (3.18), равна:

$$p=\iint_{\sigma}\varphi(x,y)d\sigma=\iint_{\sigma}\frac{d\sigma}{\pi^2\cdot(1+x^2)\cdot(1+y^2)}=\frac{1}{\pi^2}\int_0^1\frac{dy}{1+y^2}\cdot\int_0^1\frac{dx}{1+x^2}=$$

$$=\frac{1}{\pi^2}\cdot\left(arctg\,y\Big|_0^1\cdot arctg\,x\Big|_0^1\right)=\frac{1}{\pi^2}\cdot(arctg\,1-arctg\,0)\cdot(arctg\,1-arctg\,0)=\frac{1}{\pi^2}\cdot\frac{\pi}{4}\cdot\frac{\pi}{4}=\frac{1}{16}$$

2) Используя соотношение (3.19), находим функцию распределения *F(x,y)*:

$$F(x,y)=P[(X_1<x),(X_2<y)]=\int_{-\infty}^{x}dx\int_{-\infty}^{y}\varphi(x,y)dy=\int_{-\infty}^{x}dx\int_{-\infty}^{y}\frac{dy}{\pi^2(1+x^2)\cdot(1+y^2)}=$$

$$=\frac{1}{\pi^2}\int_{-\infty}^{x}\frac{dx}{1+x^2}\cdot\int_{-\infty}^{y}\frac{dy}{1+y^2}=\frac{1}{\pi^2}\cdot[arctg\,x-arctg(-\infty)]\cdot[arctg\,y-arctg(-\infty)]=$$

$$=\frac{1}{\pi^2}\cdot\left(arctg\,x+\frac{\pi}{2}\right)\cdot\left(arctg\,y+\frac{\pi}{2}\right)$$

3) Плотность распределения случайной величины X_1 находим по формуле (3.20):

$$\varphi_1(x)=\int_{-\infty}^{+\infty}\varphi(x,y)dy=\int_{-\infty}^{+\infty}\frac{dy}{\pi^2\cdot(1+x^2)\cdot(1+y^2)}=\frac{1}{\pi\cdot(1+x^2)}$$

Аналогичным образом, используя формулу (42), получим

$$\varphi_2(e)=\int_{-\infty}^{+\infty}\varphi(x,y)dx=\int_{-\infty}^{+\infty}\frac{dx}{\pi^2\cdot(1+x^2)\cdot(1+y^2)}=\frac{1}{\pi\cdot(1+y^2)}$$

Легко убедиться в том, что случайные величины X_1 и X_2 независимы, так как

$$\varphi(x,y)=\varphi_1(x)\cdot\varphi_2(y)$$
$$\longleftrightarrow$$

По определению двумерная случайная величина (X_1, X_2) распределена нормально, если плотность распределения системы величин X_1 и X_2 имеет вид:

$$\varphi(x,y)=\frac{1}{2\pi\cdot\sigma_1\cdot\sigma_2\sqrt{1-R^2}}\cdot e^{-\frac{1}{1-R^2}\cdot\left[\frac{(x-a_1)^2}{2\cdot\sigma_1^2}-R\cdot\frac{(x-a_1)\cdot(y-a_2)}{\sigma_1\cdot\sigma_2}+\frac{(y-a_2)^2}{2\cdot\sigma_2^2}\right]}$$

$$\varphi(x,y)=\frac{1}{2\pi\cdot\sigma_1\cdot\sigma_2\sqrt{1-R^2}}\cdot exp\left\{-\frac{1}{1-R^2}\cdot\left[\frac{(x-a_1)^2}{2\cdot\sigma_1^2}-R\cdot\frac{(x-a_1)\cdot(y-a_2)}{\sigma_1\cdot\sigma_2}+\frac{(y-a_2)^2}{2\cdot\sigma_2^2}\right]\right\}$$

где $\sigma_1>0$, $\sigma_2>0$, а R - некоторая постоянная. Можно показать [используя формулы (3.19) и (3.20)], что каждая из величин X_1 и X_2 распределена нормально:

$$\varphi_1(x)=\frac{1}{\sqrt{2\pi}\cdot\sigma_1}\cdot exp\left\{-\frac{(x-a_1)^2}{2\cdot\sigma_1^2}\right\}$$

$$\varphi_2(y)=\frac{1}{\sqrt{2\pi}\cdot\sigma_2}\cdot exp\left\{-\frac{(y-a_2)^2}{2\cdot\sigma_2^2}\right\}$$

На доказательстве этого факта мы не будем останавливаться. В частности, если X_1 и X_2 независимы, то

$$\varphi(x,y)=\varphi_1(x)\cdot\varphi_2(y)$$

Отсюда следует, что $R=0$, и, следовательно:

$$\varphi(x,y)=\frac{1}{2\pi\cdot\sigma_1\cdot\sigma_2}\cdot exp\left\{-\left[\frac{(x-a_1)^2}{2\cdot\sigma_1^2}+\frac{(y-a_2)^2}{2\cdot\sigma_2^2}\right]\right\}$$

Нетрудно убедиться в том, что справедливо и обратное утверждение: если $R=0$, то X_1 и X_2 — независимые случайные величины.

§ 2. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНЫХ ВЕЛИЧИН.

Оглавление.

В теории вероятности и во многих ее приложениях большое значение имеют различные числовые характеристики случайных величин. Основными из них являются математическое ожидание и дисперсия.

1. Математическое ожидание случайной величины и его свойства.

Рассмотрим сначала следующий пример. Пусть на завод поступила партия, состоящая из N подшипников. При этом:

m_1 - число подшипников с внешним диаметром x_1,

m_2 - число подшипников с внешним диаметром x_2,

. . . . . .

m_n - число подшипников с внешним диаметром x_n.

Здесь $m_1 + m_2 + \ldots + m_n = N$. Найдем среднее арифметическое значение $x_{ср}$ внешнего диаметра подшипника. Очевидно,

$$x_{ср} = \frac{m_1 x_1 + m_2 x_2 + \ldots + m_n x_n}{N} = \frac{m_1}{N} x_1 + \frac{m_2}{N} x_2 + \ldots + \frac{m_n}{N} x_n$$

Внешний диаметр вынутого наудачу подшипника можно рассматривать как случайную величину X, принимающую значения $x_1, x_2, \ldots x_n$, с соответствующими вероятностями $p_1 = \frac{m_1}{N}$, $p_2 = \frac{m_2}{N}$, ..., $p_n = \frac{m_n}{N}$, так как вероятность p_i появления подшипника с внешним диаметром x_i равна m_i/N. Таким образом, среднее арифметическое значение $x_{ср}$ внешнего диаметра подшипника можно определить с помощью соотношения

$$x_{ср} = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i$$

Пусть X - дискретная случайная величина с заданным законом распределения вероятностей $P(X = x_i) = p_i$ (такую таблицу для дискретной случайной величины мы уже приводили):

Значения X	x_1	x_2	. . .	x_n
Вероятности $P(X = x_i)$	p_1	p_2	. . .	p_n

Математическим ожиданием дискретной случайной величины X называется сумма парных произведений всех возможных значений случайной величины на соответствующие им вероятности, т.е.

$$M(X) = \sum_{i=1}^{n} x_i p_i \qquad (4.1)$$

В случае, если множество возможных значений дискретной случайной величины образует бесконечную последовательность x_1, x_2, ..., x_n, ..., то математическое ожидание этой случайной величины определяется как сумма ряда $\sum_{i=1}^{\infty} x_i p_i$, причем требуется, чтобы этот ряд абсолютно сходился.

Возвращаясь к разобранному выше примеру, мы видим, что средний диаметр подшипника равен математическому ожиданию случайной величины X - диаметру подшипника.

Математическим ожиданием непрерывной случайной величины X с плотностью распределения $\varphi(x)$ называется число, определяемое равенством

$$M(X) = \int_{-\infty}^{+\infty} x \cdot \varphi(x) dx \qquad (4.2)$$

При этом предполагается, что несобственный интеграл, стоящий в правой части равенства (4.2) существует.

Рассмотрим свойства математического ожидания. При этом ограничимся

доказательством только первых двух свойств, которое проведем для дискретных случайных величин.

1°. Математическое ожидание постоянной C равно этой постоянной.

Доказательство. Постоянную C можно рассматривать как случайную величину X, которая может принимать только одно значение C с вероятностью равной единице. Поэтому $M(X)=C\cdot 1=C$.

2°. Постоянный множитель можно выносить за знак математического ожидания, т.е. $M(kX)=k\,M(X)$.

Доказательство. Используя соотношение (4.1), имеем

$$M(kX)=\sum_{i=1}^{n} k\,x_i\,p_i = k\sum_{i=1}^{n} x_i\,p_i = k\,M(X)$$

3°. Математическое ожидание суммы нескольких случайных величин равно сумме математических ожиданий этих величин:

$$M(X_1+X_2+\ldots+X_n)=M(X_1)+M(X_2)+\ldots+M(X_n) \qquad (4.3)$$

4°. Математическое ожидание произведения двух независимых случайных величин равно произведению математических ожиданий этих величин:

$$M(X_1\cdot X_2)=M(X_1)\cdot M(X_2) \qquad (4.4)$$

Под суммой (произведением) двух случайных величин X_1 и X_2 понимают случайную величину $X=X_1+X_2$, возможные значения которой состоят из сумм (произведений) каждого возможного значения величины X_1 и каждого возможного значения величины X_2.

2. Дисперсия и ее свойства. Среднее квадратическое отклонение.

Во многих практически важных случаях существенным является вопрос о том, насколько велики отклонения $X-M(X)$ случайной величины от ее

математического ожидания.

Предварительно рассмотрим пример. Пусть две случайные величины X_1 и X_2 заданы следующими рядами распределения

Значения X_1	***-0,2***	***-0,1***	***0,1***	***0,2***
Вероятности $P(X_1)$	***0,25***	***0,25***	***0,25***	***0,25***
Значения X_2	***-50***	***-40***	***40***	***50***
Вероятности $P(X_2)$	***0,25***	***0,25***	***0,25***	***0,25***

Легко убедится в том, что математические ожидания этих величин одинаковы и равны нулю:

$$M(X_1) = (-0,2)\cdot 0,25 + (-0,1)\cdot 0,25 + 0,1\cdot 0,25 + 0,2\cdot 0,25 = 0$$

$$M(X_2) = (-50)\cdot 0,25 + (-40)\cdot 0,25 + 40\cdot 0,25 + 50\cdot 0,25 = 0$$

Однако разброс значений этих величин относительно их математического ожидания неодинаков. В первом случае значения, принимаемые случайной величиной X_1, близки к ее математическому ожиданию, а во втором случае далеки от него. Для оценки разброса (рассеяния) значений случайной величины около ее математического ожидания вводится новая числовая характеристика - ***дисперсия***.

Дисперсией $D(X)$ случайной величины X называется математическое ожидание квадрата отклонения случайной величины от ее математического ожидания:

$$D(X) = M[X - M(X)]^2 \qquad (4.5)$$

Казалось бы, естественным рассматривать не квадрат отклонения, а просто отклонение $X - M(X)$ случайной величины от ее математического ожидания. Однако математическое ожидание этого отклонения равно нулю, так как

$$M[X - M(X)] = M(X) - M[M(X)] = M(X) - M(X) = 0$$

Здесь мы воспользовались тем, что $M(X)$ постоянно, а математическое ожидание постоянной есть эта постоянная. Можно было бы принять за меру рассеяния математическое ожидание модуль отклонения случайной величины от ее математического ожидания: $M[|X - M(X)|]$. Однако, как правило, действия связанные с абсолютными величинами, приводят к громоздким вычислениям. Поэтому приняли то, что приняли.

Выведем теперь другую формулу для расчета дисперсии.

Пусть X - дискретная случайная величина, принимающая значения $x_1, x_2, \ldots, x_n$ соответственно с вероятностями $p_1, p_2, \ldots, p_n$. Очевидно, что случайная величина $[x_i - M(X)]^2$ принимает значения

$$[x_1 - M(X)]^2, [x_2 - M(X)]^2, \ldots, [x_n - M(X)]^2$$

с теми же вероятностями $p_1, p_2, \ldots, p_n$. Следовательно, согласно определению математического ожидания дискретной случайной величины, имеем

$$D(X) = M[X - M(X)]^2 = \sum_{i=1}^{n} [x_i - M(X)]^2 p_i \qquad (4.6)$$

Если же X - непрерывная случайная величина с плотностью распределения $\varphi(x)$, то по определению

$$D(X) = \int_{-\infty}^{+\infty} [x - M(X)]^2 \varphi(x) dx \qquad (4.7)$$

Принимая во внимание определение дисперсии и свойства математического ожидания, имеем

$$D(X) = M[X - M(X)]^2 = M\{X^2 - 2 \cdot X \cdot M(X) + [M(X)]^2\} =$$
$$= M(X^2) - 2M[X \cdot M(X)] + M[M(X)]^2$$

Так как $M(X)$ и $[M(X)]^2$ - постоянные, то, используя свойства математического ожидания, получим

$$M[X \cdot M(X)] = M(X) \cdot M(X) \qquad M[M(X)]^2 = [M(X)]^2$$

Следовательно,

$$D(X) = M(X^2) - 2 \cdot M(X) \cdot M(X) + [M(X)]^2$$

Откуда окончательно находим

$$D(X) = M(X^2) - [M(X)]^2 \qquad (4.8)$$

Рассмотрим теперь свойства дисперсии.

1°. ***Дисперсия постоянной равна нулю.***

Доказательство. Пусть $X = C$. По формуле (4.8) имеем

$$D(C) = M(C^2) - [M(C)]^2 = C^2 - C^2 = 0$$

так как математическое ожидание постоянной есть эта постоянная:

$$M(C) = C \qquad M(C^2) = C^2$$

2°. ***Постоянный множитель можно выносить за знак дисперсии, возводя его в квадрат:***

$$D(k \cdot X) = k^2 \cdot D(X) \qquad (4.9)$$

Доказательство. На основании соотношения (4.8), можно записать

$$D(k \cdot X) = M[(k \cdot X)^2] - [M(k \cdot X)]^2$$

Так как

$$M[(kX)^2] = M(k^2 X^2) = k^2 M(X^2)$$

и

$$[M(kX)]^2 = [k\,M(X)]^2 = k^2 [M(X)]^2$$

то

$$D(kX) = k^2 \{M(X^2) - [M(X)]^2\} = k^2 D(X)$$

3°. ***Если*** X_1 **и** X_2 ***- независимые случайные величины, то дисперсия суммы этих величин равна сумме их дисперсий:***

$$D(X_1 + X_2) = D(X_1) + D(X_2) \qquad (4.10)$$

Доказательство. По формуле (4.8) имеем

$$D(X_1 + X_2) = M[(X_1 + X_2)^2] - [M(X_1 + X_2)]^2$$

Но

$$M\left[(X_1+X_2)^2\right]=M\left(X_1^2+2X_1X_2+X_2^2\right)=M\left(X_1^2\right)+2M(X_1\cdot X_2)+M\left(X_2^2\right)$$

Так как X_1 и X_2 - независимые случайные величины, то

$$M(X_1\cdot X_2)=M(X_1)\cdot M(X_2)$$

Следовательно

$$M\left[(X_1+X_2)^2\right]=M\left(X_1^2\right)+2M(X_1)\cdot M(X_2)+M\left(X_2^2\right)$$

Далее,

$$M(X_1+X_2)=M(X_1)+M(X_2)$$

поэтому

$$[M(X_1+X_2)]^2=[M(X_1)+M(X_2)]^2=[M(X_1)]^2+2\cdot M(X_1)\cdot M(X_2)+[M(X_2)]^2$$

Таким образом

$$D(X_1+X_2)=M\left(X_1^2\right)+2\cdot M(X_1)\cdot M(X_2)+M\left(X_2^2\right)-[M(X_1)]^2-2\cdot M(X_1)\cdot M(X_2)-[M(X_2)]^2=$$

$$=\left\{M\left(X_1^2\right)-[M(X_1)]^2\right\}+\left\{M\left(X_2^2\right)-[M(X_2)]^2\right\}=D(\xi)+D(X_2)$$

Следовательно

$$D(X_1+X_2)=D(X_1)+D(X_2)$$

Замечание. Свойство 3° распространяется на любое конечное число попарно независимых случайных величин:

$$D(X_1+X_2+\ldots+X_n)=D(X_1)+D(X_2)+\ldots+D(X_n)$$

Средним квадратическим отклонением $\sigma(X)$ **случайной величины** X ***называется корень квадратный из ее дисперсии:***

$$\sigma(X)=\sqrt{D(X)} \qquad (4.11)$$

Среднее квадратическое отклонение $\sigma(X)$ имеет ту же размерность, что и случайная величина X.

Пример 4.1. Случайная величина X - число очков, выпадающих при однократном бросании игральной кости. Определить: математическое

ожидание, дисперсию и средне квадратическое отклонение.

Решение: Используя формулы (4.1), (4.6) и (4.11) соответственно получим

$$M(X)=1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}=3{,}5$$

$$D(X)=(1-3{,}5)^2\cdot\frac{1}{6}+(2-3{,}5)^2\cdot\frac{1}{6}+(3-3{,}5)^2\cdot\frac{1}{6}+(4-3{,}5)^2\cdot\frac{1}{6}+(5-3{,}5)^2\cdot\frac{1}{6}+(6-3{,}5)^2\cdot\frac{1}{6}=$$
$$=2{,}92$$

$$\sigma(X)=\sqrt{D(X)}=\sqrt{2{,}92}\approx 1{,}71$$

3. Числовые характеристики некоторых случайных величин.

Найдем теперь числовые характеристики случайных величин (математическое ожидание, дисперсию и среднее квадратическое отклонение) случайных величин, рассмотренных выше.

1. ***Распределение Бернулли***. Здесь случайная величина X - число наступления события A при одном испытании, причем $P(A)=p$. Найти математическое ожидание и дисперсию для этого распределения.

Величина X принимает два значения 0 и 1 соответственно с вероятностями $q=1-p$ и p. Поэтому по формулам (4.1) и (4.6) находим

$$M(X)=0\cdot(1-p)+1\cdot p=p$$

$$D(X)=(0-p)^2(1-p)+(1-p)^2 p=p\cdot(1-p)=p\cdot q$$

2. ***Биноминальный закон распределения***. Определяется формулой Бернулли: $P_n(m)=\dfrac{n!}{m!\ (n-m)!}p^m\cdot q^{n-m}$, где p - постоянная вероятность появления события A в данном конкретном опыте.

Пусть X - случайная величина, принимающая значения 1 или 0 в зависимости от того, происходит или не происходит событие A в i-м опыте. Тогда $m=X_1+X_2+\ldots+X_n$. Ясно, что X попарно независимы. Из результата примера 2 следует, что $M(X)=p$, $D(X)=p\cdot q$ для любого i. На основании свойства 3° для математического ожидания и дисперсии имеем

$$M(m)=M(X_1+X_2+\ldots+X_n)=M(X_1)+M(X_2)+\ldots+M(X_n)=n\cdot p$$

$$D(m)=D(X_1+X_2+\ldots+X_n)=D(X_1)+D(X_2)+\ldots+D(X_n)=n\cdot p\cdot q$$

$$\sigma(m)=\sqrt{npq}$$

3. Пусть X - случайная величина, распределенная по ***закону Пуассона:*** $p(k)=P(X=k)=\dfrac{\lambda^k e^{-\lambda}}{k!}$ $(k=0,1,\ldots n,\ldots)$. Найти $M(X)$ и $D(X)$.

Используя соотношение (4.1), получим

$$M(X)=\sum_{k=0}^{\infty}k\cdot\frac{\lambda^k e^{-\lambda}}{k!}=e^{-\lambda}\cdot\lambda\sum_{k=0}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=e^{-\lambda}\cdot\lambda\cdot e^{\lambda}=\lambda$$

Так как

$$\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=1+\frac{\lambda}{1!}+\frac{\lambda^2}{1!}+\ldots+\frac{\lambda^k}{k!}+\ldots=e^{\lambda}$$

Найдем теперь выражение для дисперсии закона Пуассона

$$D(X)=M(X^2)-\left[M(X)^2\right]=M(X^2)-\lambda^2$$

$$M(X^2)=\sum_{k=0}^{\infty}k^2\cdot\frac{\lambda^k e^{-\lambda}}{k!}=e^{-\lambda}\cdot\lambda\sum_{k=0}^{\infty}\frac{k\cdot\lambda^{k-1}}{(k-1)!}=e^{-\lambda}\cdot\lambda\sum_{k=0}^{\infty}\frac{(k+1-1)\cdot\lambda^{k-1}}{(k-1)!}=$$

$$=e^{-\lambda}\cdot\lambda\sum_{k=0}^{\infty}\frac{(k-1)\cdot\lambda^{k-1}}{(k-1)!}+e^{-\lambda}\cdot\lambda\sum_{k=0}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=e^{-\lambda}\cdot\lambda^2\cdot\sum_{k=0}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+e^{-\lambda}\cdot\lambda\cdot e^{\lambda}=$$

$$=e^{-\lambda}\cdot\lambda^2\cdot e^{\lambda}+\lambda=\lambda^2+\lambda$$

Следовательно,

$$D(X)=M(X^2)-\left[M(X)^2\right]=M(X^2)-\lambda^2=\lambda^2+\lambda-\lambda^2=\lambda$$

4. Пусть теперь X - случайная величина, имеющая ***равномерное распределение*** с плотностью $\varphi(x)=\begin{cases}0, & \text{если } x<a\\ \dfrac{1}{b-a}, & \text{если } a<x<b\\ 0, & \text{если } x>b\end{cases}$

Найдем математическое ожидание, дисперсию и средне квадратическое отклонение этой случайной величины.

По формулам (4.2), (4.7) и (4.11) находим

$$M(X)=\int_{-\infty}^{+\infty} x\cdot\varphi(x)\,dx=\int_{a}^{b} x\cdot\frac{1}{b-a}dx=\frac{a+b}{2}$$

$$D(X)=\int_{-\infty}^{+\infty}\left(x-\frac{b+a}{2}\right)^2\varphi(x)dx=\int_{a}^{b}\left(x-\frac{b+a}{2}\right)^2\cdot\frac{1}{b-a}dx=\frac{(b-a)^2}{12}$$

$$\sigma(X)=\sqrt{D(X)}=\sqrt{\frac{(b-a)^2}{12}}=\frac{b-a}{2\cdot\sqrt{3}}=\sqrt{3}\cdot\frac{b-a}{6}$$

5. Сейчас можно выяснить смысл параметров a и σ ***нормального закона*** распределения случайной величины.

Пусть X - нормально распределенная случайная величина, с параметрами a и σ. Найдем $M(X)$ и $D(X)$.

Так как $\varphi(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-a)^2/(2\sigma^2)}$, то по формуле (4.2) находим

$$M(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} x\,e^{-(x-a)^2/(2\sigma^2)}dx$$

Проведем в интеграле замену переменной, полагая $\frac{x-a}{\sigma}=z$. Тогда $x=a+\sigma z$, $dx=\sigma\,dz$. Следовательно,

$$M(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} x\,e^{-(x-a)^2/(2\sigma^2)}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}(a+\sigma z)\cdot e^{-z^2/2}dz=$$

$$=\frac{a}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-z^2/2}dz+\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-z^2/2}z\,dz$$

Но

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-z^2/2}dz=1$$

см. формулу (3.7). Далее, так как функция $e^{-z^2/2}z$ нечетная, то по свойству нечетных функций $\int_{-\infty}^{+\infty} e^{-z^2/2}z\,dz=0$. Следовательно, $M(X)=a$.

Дисперсию находим по формуле (4.7)

$$D(\xi)=\int_{-\infty}^{+\infty}[x-M(\xi)]^2\varphi(x)dx=\int_{-\infty}^{+\infty}(x-a)^2\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-a)^2/(2\sigma^2)}dx=\sigma^2$$

(вычисление интеграла не приводим).

Итак, $D(X)=\sigma^2$ $\sigma(X)=\sqrt{D(X)}=\sigma$.

Таким образом, параметры a и σ для нормально распределенной случайной величины имеют простой вероятностный смысл: a есть математическое ожидание, σ - среднее квадратическое отклонение.

4. Линейные функции случайных величин.

Пусть X - нормально распределенная случайная величина с параметрами $M(X)=a$ и $\sigma(X)=\sigma$. Тогда, ***если A и B - постоянные, то случайная величина*** $\eta=A+B\cdot X$, ***линейно зависящая от*** X, ***также нормально распределена, причем*** *

$$M(\eta)=A+B\cdot a, \qquad D(\eta)=B^2\sigma^2$$

Докажем это утверждение. Пусть для простоты $B>0$. Оценим вероятность неравенств $y_1<\eta<y_2$. Ясно, что эти неравенства равносильны неравенствам $y_1<A+B\cdot X<y_2$, т.е.

$$\frac{y_1-A}{B}<X<\frac{y_2-A}{B}$$

Поэтому $$P(y_1<\eta<y_2)=P\left(\frac{y_1-A}{B}<X<\frac{y_2-A}{B}\right)$$

Так как величина X распределена нормально, то

$$P\left(\frac{y_1-A}{B}<X<\frac{y_2-A}{B}\right)=\frac{1}{\sigma\sqrt{2\pi}}\int_{(y_1-A)/B}^{(y_2-A)/B}e^{-(x-a)^2/(2\sigma^2)}dx$$

Проведем в этом интеграле замену переменной, полагая $x=\frac{y-A}{B}$.

Тогда $dx=\frac{dy}{B}$ и, следовательно,

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{(y_1-A)/B}^{(y_2-A)/B}e^{-(x-a)^2/(2\sigma^2)}dx=\frac{1}{\sigma B\sqrt{2\pi}}\int_{y_1}^{y_2}e^{-(y-A-aB)^2/(2\sigma^2B^2)}dy$$

Итак,

$$P(y_1 < \eta < y_2) = \frac{1}{\sigma B\sqrt{2\pi}} \int_{y_1}^{y_2} e^{-(y - A - aB)^2/(2\sigma^2 B^2)}\, dy$$

Это равенство показывает, что случайная величина η имеет нормальное распределение, причем $M(\eta) = A + Ba$ и $D(\eta) = \sigma^2 B^2$.

Имеет место и более общее утверждение. Пусть $\lambda_1, \lambda_2, \ldots, \lambda_n$ - постоянные, а $X_1, X_2, \ldots X_n$ - нормально распределенные попарно независимые случайные величины, причем $M(X_i) = a \quad D(X_i) = \sigma_i^2$.

Тогда случайная величина $\eta = \lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_n X_n$

также имеет нормальное распределение, причем

$$M(\eta) = \lambda_1 a_1 + \lambda_2 a_2 + \ldots + \lambda_n a_n$$

$$D(\eta) = \lambda_1^2 \sigma_1^2 + \lambda_2^2 \sigma_2^2 + \ldots + \lambda_n^2 \sigma_n^2$$

В частности, если $M(X_i) = a \quad D(X_i) = \sigma^2$ при любом i, то случайная величина $\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$ распределена нормально, причем

$M(\overline{X}) = a$, $D(\overline{X}) = \sigma^2 / n$, $\sigma(\overline{X}) = \sqrt{D(\overline{X})} = \sigma / \sqrt{n}$.

* Это утверждение можно получить просто из свойств математического ожидания и дисперсии.

§ 3. ПРИМЕНЕНИЕ ТЕОРИИ ВЕРОЯТНОСТИ К СТАТИСТИКЕ.

Оглавление.

1. Основные понятия.

Математическая статистика - это раздел математики, в котором изучаются методы обработки и анализа экспериментальных данных, полученных в результате наблюдений над массовыми случайными событиями, явлениями.

Наблюдения, проводимые над объектами, могут охватывать всех членов изучаемой совокупности без исключения и могут ограничиваться обследованиями лишь некоторой части членов данной совокупности. Первое наблюдение называется сплошным или полным, второе частичным или ***выборочным*.**

Естественно, что наиболее полную информацию дает сплошное наблюдение, однако к нему прибегают далеко не всегда. Во-первых, сплошное наблюдение очень трудоемко, а во-вторых, часто бывает практически невозможно или даже нецелесообразно. Поэтому в подавляющем большинстве случаев прибегают к выборочному исследованию.

Совокупность, из которой некоторым образом отбирается часть ее членов для совместного изучения, называется **генеральной совокупностью**, а отобранная тем или иным способом часть генеральной совокупности - выборочная совокупность или **выборка**.

Объем генеральной совокупности N теоретически ничем неограничен

$N \to \infty$, на практике же он всегда ограничен.

Объем выборки n может быть большим или малым, но он не может быть меньше двух.

Отбор в выборку можно проводить случайным способом (по способу жеребьевки или лотереи). Либо планово, в зависимости от задачи и организации обследования. Для того, чтобы выборка была представительной, необходимо обращать внимание на размах варьирования признака и согласовывать с ним объем выборки.

2. Определение неизвестной функции распределения.

Итак, мы сделали выборку. Разобьем диапазон наблюдаемых значений X на интервалы $]x_0, x_1[$, $]x_1, x_2[$, ... $]x_{k-1}, x_k[$ одинаковой длины Δx. Для оценки необходимого числа интервалов k можно использовать следующие формулы:

$$k = 1 + 3{,}32 \cdot \lg n \qquad \textit{или} \qquad k = 5 \cdot \lg n . \tag{5.1}$$

Далее пусть m_i - число наблюдаемых значений X, попавших в ***i*-ый** интервал. Разделив m_i на общее число наблюдений n, получим частоту p_i^*, соответствующую ***i*-ому** интервалу: $p_i^* = \frac{m_i}{n}$, причем $\sum_{i=1}^{k} p_i^* = \sum_{i=1}^{k} \frac{m_i}{n} = 1$.

Составим следующую таблицу:

Номер интервала	Интервал	m_i	p_i^*
1	$]x_0, x$	m_1	p_1^*
2	$]x_1, x$	m_2	p_2^*
...	...	...	...
k	$]x_{k-1}$	m_k	p_k^*

которая называется ***статистическим рядом. Эмпирической*** (или

статистической) ***функцией распределения*** случайной величины X называется частота события, заключающегося в том, что величина X в результате опыта примет значение, меньшее x:

$$F^*(x) = P^*(X < x)$$

На практике достаточно найти значения статистической функции распределения $F^*(x)$ в точках $x_0, x_2, \dots x_k$, которые являются границами интервалов статистического ряда:

$$\begin{cases} F^*(x_0) = P^*(\xi < x_0) = 0 \\ F^*(x_1) = P^*(\xi < x_1) = \dfrac{m_1}{n} = p_1^* \\ F^*(x_2) = P^*(\xi < x_2) = \dfrac{m_1 + m_2}{n} = p_1^* + p_2^* \\ \dots\dots\dots\dots\dots \\ F^*(x_k) = P^*(\xi < x_k) = \dfrac{m_1 + m_2 + \dots + m_k}{n} = p_1^* + p_2^* + \dots + p_k^* = 1 \end{cases} \qquad (5.2)$$

Следует заметить, что $F^*(x) = 0$ при $x < x_0$ и $F^*(x) = 1$ при $x > x_k$. Построив точки $M_i\,]x_i ; F^*(x_i)[$ и соединив их плавной кривой, получим приближенный график эмпирической функции распределения (рис. 5.1). Используя закон больших чисел Бернулли, можно доказать, что при достаточно большом числе испытаний n с вероятностью, близкой к единице, эмпирическая функция распределения $F^*(x)$ отличается сколь угодно мало от неизвестной нам функции распределения $F(x)$ случайной величины X.

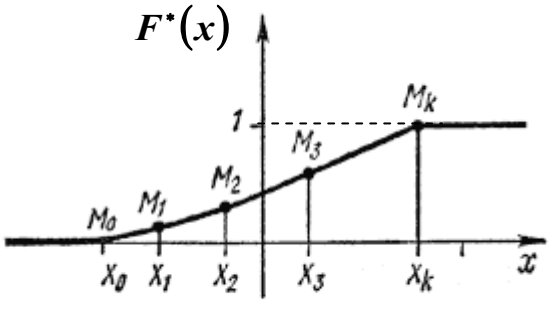

Рис. 5.1

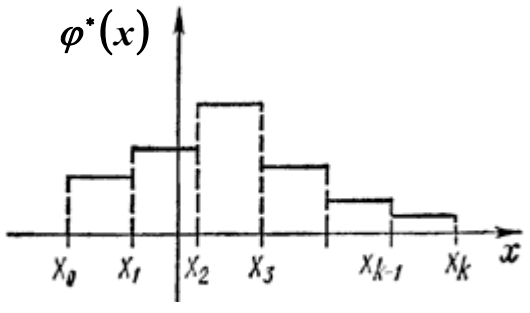

Рис. 5.2

Часто вместо построения графика эмпирической функции распределения поступают следующим образом. На оси абсцисс откладывают интервалы $]x_0, x_1[$, $]x_1, x_2[$,.... $]x_{k-1}, x_k[$. На каждом интервале строят прямоугольник, площадь которого равна частоте p_i^*, соответствующей данному интервалу. Высота h_i этого прямоугольника равна $h_i = \frac{p_i^*}{\Delta x}$, где Δx - длинна каждого из интервалов. Ясно, что сумма площадей всех построенных прямоугольников равна единице.

Рассмотрим функцию $y = \varphi^*(x)$, которая в интервале $]x_{i-1}, x_i[$ постоянна и равна h_i. График этой функции называется ***гистограммой***. Он представляет собой ступенчатую линию (рис. 5.2). С помощью закона больших чисел Бернулли можно доказать, что при малых Δx и больших n с практической достоверностью $\varphi^*(x)$ как угодно мало отличается от плотности распределения $\varphi(x)$ непрерывной случайной величины X.

Таким образом на практике определяется вид неизвестной функции распределения случайной величины.

3. Определение неизвестных параметров распределения.

Таким образом мы получили гистограмму, которая дает наглядность. Наглядность представленных результатов позволяет сделать различные заключения, суждения об исследуемом объекте.

Однако на этом обычно не останавливаются, а идут дальше, анализируя данные на проверку определенных предположений относительно возможных механизмов изучаемых процессов или явлений.

Несмотря на то, что данных в каждом обследовании сравнительно немного, мы бы хотели, чтобы результаты анализа достаточно хорошо описывали бы все реально существующее или мыслимое множество (т.е. генеральную совокупность).

Для этого делают некоторые предположения о том, как вычисленные на основе экспериментальных данных (выборке) показатели соотносятся с параметрами генеральной совокупности.

Решение этой задачи составляет главную часть любого анализа экспериментальных данных и тесно связано с использованием ряда теоретических распределений, рассмотренных выше.

Широкое использование в статистических выводах нормального распределения имеет под собой как эмпирическое, так и теоретическое обоснование.

Во-первых, практика показывает, что во многих случаях нормальное распределение действительно является довольно точным представлением экспериментальных данных.

Во-вторых, теоретически показано, что средние значения интервалов гистограмм распределены по закону, близкому к нормальному.

Однако следует четко представлять, что нормальное распределение - это лишь чисто математический инструмент и совсем необязательно, чтобы реальные экспериментальные данные точно описывались нормальным распределением. Хотя во многих случаях, допуская небольшую ошибку, можно говорить, что данные распределены нормально.

Ряд показателей, такие как среднее, дисперсия и т.д., характеризуют выборку и называются статистиками. Такие же показатели, но относящиеся к генеральной совокупности в целом, называются параметрами. Таким образом, можно сказать, что статистики служат для оценки параметров.

Генеральной средней $x_Г$ называется среднее арифметическое значений $x_1, x_2, \ldots x_N$ генеральной совокупности объема N:

$$x_Г = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Выборочной средней $x_В$ называется среднее арифметическое выборки $x_1, x_2, \ldots x_n$ объема n:

$$x_B = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \tag{5.3}$$

или

$$x_B = \frac{1}{n}\sum_{i=1}^{n} n_i \cdot x_i \tag{5.4}$$

если выборка имеет вид таблицы.

Выборочную среднюю принимают в качестве оценки генеральной средней.

Генеральной дисперсией $D_Г$ называется среднее арифметическое квадратов отклонения значений генеральной совокупности $x_1, x_2, \ldots x_N$ от их среднего значения $x_Г$:

$$D_Г = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - x_Г\right)^2$$

Генеральным средним квадратическим отклонением $\sigma_Г$ называется корень квадратный из генеральной дисперсии: $\sigma_Г = \sqrt{D_Г}$.

Выборочной дисперсией D_B называется среднее арифметическое квадратов отклонения значений выборки $x_1, x_2, \ldots x_n$ от их среднего значения $x_в$:

$$D_B = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - x_B\right)^2 \qquad \textit{или} \qquad D_B = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot \left(x_i - x_B\right)^2$$

Выборочное среднее квадратическое отклонение σ_B определяется как $\sigma_B = \sqrt{D_B}$.

Для лучшего совпадения с результатами экспериментов, вводят понятие эмпирической (или исправленной) дисперсии s^2:

$$s^2 = \frac{n}{n-1} D_B \,, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i \cdot \left(x_i - x_B\right)^2$$

Для оценки генерального среднего квадратического отклонения служит исправленное среднее квадратическое отклонение, или эмпирический стандарт s:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} n_i \cdot (x_i - x_B)^2} \qquad (5.5)$$

В случае, когда все значения выборки $x_1, x_2, \ldots x_n$ различны, т.е. $n_i = 1$, $k = n$, формулы для s^2 и s принимают вид:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - x_B)^2} \qquad (5.6)$$

4. Доверительный интервал. Доверительная вероятность.

Различные статистики, получаемые результате вычислений, представляют собой точечные оценки соответствующих параметров генеральной совокупности.

Если из генеральной совокупности извлечь некоторое количество выборок и для каждой из них найти интересующие нас статистики, то вычисленные значения будут представлять собой случайные величины, имеющие некоторый разброс вокруг оцениваемого параметра.

Но, как правило, в результате эксперимента в распоряжении исследователя имеется одна выборка. Поэтому значительный интерес представляет получение интервальной оценки, т.е. некоторого интервала, внутри которого, как можно предположить, лежит истинное значение параметра.

Вероятности, признанные достаточными для уверенных суждениях о параметрах генеральной совокупности на основании статистик, называются доверительными.

Для примера рассмотрим x_B как оценку параметра a .

Известно, что если выборки извлекаются из генеральной совокупности с параметрами:

$$M(X) = a \qquad D(X) = \sigma^2$$

то распределение выборочных средних $x_в$ будет иметь среднее, равное a, дисперсию $s_X = \sigma^2 / n$, среднее квадратическое $\sigma_X = \sigma / \sqrt{n}$, где n - объем выборки и будет приближаться к нормальному.

Для такого распределения, как известно, $68\,\%$ наблюдений лежит в интервале $x_в \pm \sigma / \sqrt{n}$, $95\,\%$ в интервале $x_в \pm 2 \cdot \sigma / \sqrt{n}$ и $99\,\%$ в интервале $x_в \pm 3 \cdot \sigma / \sqrt{n}$.

$$P\left(x_в - \frac{\sigma t}{\sqrt{n}} < a < x_в + \frac{\sigma t}{\sqrt{n}} \right) = \gamma \qquad (5.7)$$

где $\gamma = 2\Phi(t)$.

С надежностью γ доверительный интервал $\left(\bar{x} - t\frac{\sigma}{\sqrt{n}} < a < \bar{x} - t\frac{\sigma}{\sqrt{n}} \right)$ покрывает неизвестный параметр a с точностью $\delta = t\frac{\sigma}{\sqrt{n}}$. Здесь мы задаемся надежностью γ, а зная γ по таблицам для функции Лапласа находим параметр t и далее - доверительный интервал.

Но истинное значение параметра генеральной совокупности σ нам неизвестно. Поэтому на практике вместо параметра σ используют выборочное среднее квадратическое отклонение s. То есть доверительный интервал определяется выражением

$$\left(x_в - \frac{s \cdot t}{\sqrt{n}},\ x_в + \frac{s \cdot t}{\sqrt{n}} \right) \qquad (5.8)$$

Но здесь параметр t уже параметр распределения Стьюдента, который находится по соответствующим таблицам при данных n и γ, где γ - задаваемая надежность. Этот интервал покрывает неизвестный параметр a с надежностью γ, где $x_в$ и s находятся по формулам (5,3), (5.4) и (5.5), (5.6) соответственно.

Пример. Найти доверительный интервал для оценки математического ожидания a нормальной случайной величины с надежностью $\gamma = 0{,}95$, зная

выборочную среднюю $x_в = 75{,}15$, объем выборки $n = 64$, среднее квадратическое отклонение $\sigma = 8$.

Решение. Имеем $2\Phi(t) = 0{,}95$. Отсюда $\Phi(t) = 0{,}475$. По таблице значений функции Лапласа находим $t = 1{,}96$. Отсюда

$$\left(x_в - \frac{\sigma t}{\sqrt{n}},\ x_в + \frac{\sigma t}{\sqrt{n}} \right) \Rightarrow \left(75{,}15 - \frac{8 \cdot 1{,}96}{\sqrt{64}} < a < 75{,}15 + \frac{8 \cdot 1{,}96}{\sqrt{64}} \right)$$

$$(75{,}15 - 1{,}96 < a < 75{,}15 + 1{,}96) \Rightarrow (73{,}19 < a < 77{,}11)$$

5. Применение критерия Стьюдента для сравнения генеральных совокупностей.

Например, нам надо оценить эффективность действия рекламы какого-то товара. До запуска рекламы продажа товара по неделям (в шт.) имела следующий вид:

$$70, 78, 60, 80, 60, 60, 68;\quad n_1 = 7;\quad x_{в1} = 68;\quad s_1^2 = 440;\quad s_1 \approx 21$$

После выпуска рекламы продажа этого же товара по неделям стала иметь вид:

$$80, 75, 62, 70, 68, 71;\quad n_2 = 6;\quad x_{в2} = 71;\quad s_2^2 = 188;\quad s_2 \approx 13{,}7$$

Следовательно, доверительный интервал с надежностью $95\,\%$ для первой выборки равен

$$68 \pm \frac{2 \cdot 21}{\sqrt{7}} = 68 \pm 16$$

А для второй

$$71 \pm \frac{2 \cdot 13{,}7}{\sqrt{6}} = 71 \pm 11$$

Таким образом, если по средним мы можем сделать положительный вывод о влиянии рекламы товара, то по доверительным интервалам мы вправе сомневаться: уж очень велики интервалы и они значительно перекрывают друг друга (см. рис. 5.3).

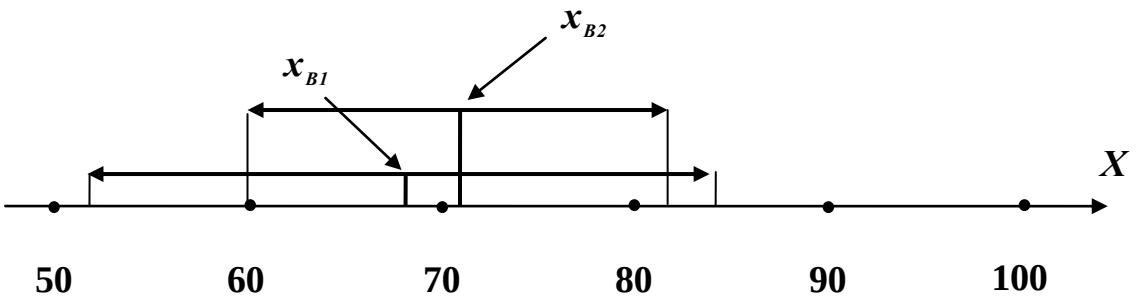

Рис. 5.3

Однако нам необходимо со всей определенностью истолковать результаты эксперимента.

Мы можем высказать два предположения (статистические гипотезы).

1. Нулевая гипотеза. Между генеральными совокупностями с параметрами a_1 и a_2, σ_1 и σ_2 разница равна нулю, т.е. $a_1 - a_2$. Следовательно, разница между выборочными средними $d = x_{B1} - x_{B2}$ возникла случайно, в процессе группировки данных.

2. Альтернативная гипотеза, т.е. противоположная.

Для проверки этих гипотез существуют специальные параметры, которые табулированы и приводятся в соответствующих справочниках.

В частности, если сравниваемые генеральные совокупности имеют нормальный закон распределения, то сравнение выборочных средних проводят с помощью ***t - критерия*** или критерия Стьюдента:

$$t = \frac{(x_{B1} - x_{B2}) - (a_1 - a_2)}{S_d}$$

$$S_d = \sqrt{\frac{\sum (x_i - x_{B1})^2 + \sum (x_i - x_{B2})^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}}.$$

Согласно нулевой гипотезе $a_1 = a_2$ $a_1 - a_2$, отсюда:

$$t = \frac{x_{B1} - x_{B2}}{S_d} = \frac{d}{S_d} \qquad (5.9)$$

Нулевая гипотеза (разницы нет) отвергается, если $t_Ф > t_{St}$ для заданной надежности и числа (степеней свободы) $k = n_1 + n_2 - 2$. Здесь $t_Ф$ - фактический коэффициент Стьюдента, найденный по формуле (5.9), а t_{St} - теоретический коэффициент, найденный по специальным таблицам.

Для нашего примера $d = 71 - 68 = 3$, $S_d = \sqrt{\frac{440 + 188}{7 + 6 - 2} \cdot \frac{7 + 6}{7 \cdot 6}} = 4{,}2$.

Следовательно, $t_Ф = \frac{3}{4{,}2} = 0{,}71$. По таблицам, для надежности $95\,\%$ и числа $k = 7 + 6 - 2 = 11$, находим $t_{St} = 2{,}2$. Итак, $t_Ф < t_{St}$ и нулевая гипотеза сохраняется: разница между результатами опыта и контроля оказалась статистически недостоверной.

Таблица ***t - критерия*** Стьюдента.

k	Уровни надежности		
	95 %	99 %	99,9 %
7	2,37	3,50	5,51
8	2,31	3,36	5,04
9	2,26	3,25	4,78
10	2,23	3,17	4,59

6. Элементы теории корреляции.

Между различного рода признаками, случайными величинами практически всегда существует взаимосвязь. Только иногда эту связь мы замечаем, но в большинстве случаев эти взаимосвязи ускользают от нашего внимания.

В одних случаях получается функциональная связь, когда между признаками x и y существует однозначная зависимость: $y = f(x)$. Например $S = \pi \cdot R^2$, $V = \frac{4}{3} \cdot \pi \cdot R^3$ и т.д.

В других случаях получается корреляционная зависимость, когда одному значению признака X соответствуют несколько значений признака Y. То есть здесь мы имеем дело со статистической связью. Например, связь между ростом человека и его весом, связь между стажем работника и качеством его труда и т.д.

Корреляционная связь между признаками может быть линейной и нелинейной, положительной и отрицательной. Задача корреляционного анализа сводится к установлению формы и направления связи между признаками, измерению ее тесноты и к оценке достоверности выборочных коэффициентов корреляции.

Корреляционным моментом μ_{XY} случайных величин X и Y называют математическое ожидание произведения отклонений этих величин от своих математических ожиданий:

$$\mu_{XY} = M((X - M(X)) \cdot (Y - M(Y)))$$

Корреляционный момент служит для характеристики связи между величинами X и Y.

Корреляционный момент равен нулю, если X и Y независимы, следовательно, если корреляционный момент не равен нулю, то X и Y — в какой-то степени зависимые случайные величины.

Теорема 1. Корреляционный момент двух независимых случайных величин X и Y равен нулю.

Доказательство: т.к. X и Y — независимые случайные величины, то их отклонения от своих математических ожиданий $X - M(X)$ и $Y - M(Y)$ также независимы. Пользуясь свойствами математического ожидания (математическое ожидание произведения независимых случайных величин

равно произведению математических ожиданий сомножителей) и отклонения (математическое ожидание отклонения равно нулю), получим

$$\mu_{XY} = M\left(\left(X - M(X)\right)\cdot\left(Y - M(Y)\right)\right) = M\left(X - M(X)\right)\cdot M\left(Y - M(Y)\right) = 0\cdot 0 = 0$$

Из определения корреляционного момента следует, что он имеет размерность, равную произведению размерностей величин X и Y, т.е. величина корреляционного момента зависит от единиц измерения случайных величин. Поэтому для одних и тех же двух величин величина корреляционного момента имеет различные значения в зависимости от того, в каких единицах были измерены величины.

Такая особенность корреляционного момента является недостатком этой числовой характеристики, т.к. сравнение корреляционных моментов различных систем случайных величин становится затруднительным. Для того чтобы устранить этот недостаток, вводят новую числовую характеристику— коэффициент корреляции r_{XY}.

Коэффициентом корреляции случайных величин X и Y называют отношение корреляционного момента к произведению средних квадратических отклонений этих величин:

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \cdot \sigma_y}$$

Так как размерность μ_{xy} равна произведению размерностей величин X и Y, σ_x имеет размерность величины X, σ_y имеет размерность величины Y, то r_{xy} — безразмерная величина.

Таким образом, величина коэффициента корреляции не зависит от выбора единиц измерения случайных величин. В этом и состоит преимущество коэффициента корреляции перед корреляционным моментом.

Очевидно, коэффициент корреляции независимых случайных величин равен нулю (т.к. $\mu_{xy} = 0$).

Абсолютная величина коэффициента корреляции не превышает единицы: $|r_{xy}| \le 1$

На практике мы имеем дело с выборками, а не с генеральными совокупностями. Поэтому на практике рассчитывают выборочный коэффициент корреляции, который может быть достоверным или нет. Выборочный коэффициент корреляции рассчитывается по следующей формуле:

$$r_{XY} = \frac{\sum_{i=}^{n}(x_i - x_{cp})\cdot(y_i - y_{cp})}{\sqrt{\sum_{i=1}^{n}(x_i - x_{cp})^2 \cdot \sum_{i=1}^{n}(y_i - y_{cp})^2}} \qquad (5.10)$$

Коэффициент корреляции удобный показатель связи, получивший широкое применение в практике. Это отвлеченное число, лежащее в пределах от -1 до $+1$. При независимом варьировании признаков, когда связь между ними отсутствует, $r = 0$. При $r > 0$ существует положительная связь между признаками (с ростом x растет и y). При $r < 0$ - отрицательная связь - с ростом признака x признак y уменьшается. Чем больше r по модулю, тем теснее связь между признаками. При $r = 1$ между признаками существует функциональная связь.

Лишь один недостаток имеется у этого ценного показателя - он способен характеризовать лишь линейный связи. При наличии нелинейной связи между коррелирующими признаками следует использовать другие показатели.

Выборочный коэффициент корреляции служит оценкой генерального параметра $r_Г$, и, как случайная величина, сопровождается ошибками. Поэтому здесь также проверяется гипотеза о значимости выборочного коэффициента корреляции.

Пусть двумерная генеральная совокупность (X, Y) распределена нормально. Из этой совокупности извлечены выборки объемом n и по ним найден выборочный коэффициент корреляции $r_В$, который оказался

отличным от нуля. Так как выборки отобраны случайно, еще нельзя заключить, что коэффициент корреляции генеральной совокупности $r_Г$ также отличен от нуля. А, поскольку нас интересует именно этот коэффициент, возникает необходимость при заданном уровне значимости α проверить нулевую гипотезу $H_0: \rho = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $H_1: r_Г \neq 0$.

Если нулевая гипотеза отвергается, значит, выборочный коэффициент корреляции значимо отличается от нуля (кратко говоря, значим), а X и Y коррелированны, т. е. связаны линейной зависимостью.

Если же нулевая гипотеза будет принята, значит, выборочный коэффициент корреляции является незначимым, а X и Y некоррелированные, т. е. не связаны линейной зависимостью.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$t = r_В \cdot \frac{\sqrt{n-2}}{\sqrt{1-r_В^2}}$$

Величина t при справедливости нулевой гипотезы имеет распределение Стьюдента с $k = n-2$ степенями свободы.

Обозначим значение критерия, вычисленное по данным наблюдений, через $t_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: r_Г = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $H_1: r_Г \neq 0$, надо вычислить наблюдаемое значение критерия:

$$t_{набл} = r_В \cdot \frac{\sqrt{n-2}}{\sqrt{1-r_В^2}} \qquad (5.11)$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости и числу степеней свободы $k = n-2$ найти критическую точку $t_{кр}(\alpha, k)$.

Если $|t_{набл}| < t_{кр}(\alpha, k)$ — нет оснований отвергнуть нулевую гипотезу, если $|t_{набл}| > t_{кр}(\alpha, k)$ - то ее отвергают.

В то время как задача корреляционного анализа - установить, являются ли данные случайные величины взаимосвязанными, цель регрессионного анализа - описать эту связь аналитической зависимостью, т.е. с помощью уравнения. Мы рассмотрим самый несложный случай, когда связь между точками на графике может быть представлена прямой линией. Уравнение этой прямой линии $Y = a \cdot X + b$, где

$$a = Y_{ср} - b \cdot X_{ср}, \qquad b = \frac{\sum_{i=1}^{n}(X_i - X_{ср}) \cdot (Y_i - Y_{ср})}{\sum_{i=1}^{n}(X_i - X_{ср})^2} \qquad (5.12)$$

Зная уравнение прямой, мы можем находить значение функции по значению аргумента в тех точках, где значение X известно, а Y - нет. Эти оценки бывают очень нужны, но они должны использоваться осторожно, особенно, если связь между величинами не слишком тесная. Отметим также, что из сопоставления формул для b и r видно, что коэффициент не дает значение наклона прямой, а лишь показывает сам факт наличия связи.

7. Проверка гипотезы о нормальном распределении генеральной совокупности. Критерий согласия Пирсона.

Ранее предполагалось, что закон распределения генеральной совокупности известен. Если же он неизвестен, но есть основания предположить, что он имеет определенный вид (назовем его A), то проверяют нулевую гипотезу: генеральная совокупность распределена по закону A.

Проверка гипотезы о предполагаемом законе неизвестного распределения производится так же, как и проверка гипотезы о параметрах распределения, т. е. при помощи специально подобранной случайной величины — критерия согласия.

Критерием согласия c^2 называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Имеется несколько критериев согласия, но мы ограничимся описанием применения критерия Пирсона к проверке гипотезы о нормальном распределении генеральной совокупности (критерий аналогично применяется и для др. распределений). Для этого будем сравнивать эмпирические (наблюдаемые) и теоретические (вычисленные в предположении нормального распределения) частоты.

Обычно эмпирические и теоретические частоты различаются. Возможно, что расхождение случайно (незначимо) и объясняется либо малым числом наблюдений, либо способом их группировки, либо другими причинами. Возможно, что расхождение частот неслучайно (значимо) и объясняется тем, что теоретические частоты вычислены исходя из неверной гипотезы о нормальном распределении генеральной совокупности.

Критерий Пирсона отвечает на вопрос «***Случайно ли расхождение частот?***». Правда, как и любой критерий, он не доказывает справедливость гипотезы, а лишь устанавливает на принятом уровне значимости ее согласие или несогласие с данными наблюдений.

Итак, пусть по выборке объема n получено эмпирическое распределение: варианты - x_i: $x_1, x_2, \dots x_s$, эмпирические частоты - n_i: $n_1, n_2, \dots n_s$.

Допустим, что в предположении нормального распределения генеральной совокупности вычислены теоретические частоты n'_i. При уровне значимости α требуется проверить нулевую гипотезу: генеральная совокупность распределена нормально.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$c^2 = \sum \frac{(n_i - n'_i)^2}{n'_i}.$$

Эта величина случайная, т.к. в различных опытах она принимает различные, заранее не известные значения. Ясно, что, чем меньше различаются эмпирические и теоретические частоты, тем меньше величина критерия, и, следовательно, он в известной степени характеризует близость эмпирического и теоретического распределений.

Стоит заметить, что возведение в квадрат разностей частот устраняет возможность взаимного погашения положительных и отрицательных разностей. Делением на n_i' достигают уменьшения каждого из слагаемых – иначе сумма была бы настолько велика, что приводила бы к отклонению нулевой гипотезы даже тогда, когда она справедлива.

Доказано, что при $n \to \infty$ закон распределения случайной величины c^2 независимо от того, к какому закону распределения подчинена генеральная совокупность, стремится к закону распределения χ^2 с k степенями свободы. Поэтому случайная величина c^2 обозначена через χ^2, а сам критерий называют критерием согласия «***хи квадрат***».

Число степеней свободы находят по равенству $k = s - r - 1$, где s — число групп выборки; r — число параметров предполагаемого распределения, которые оценены по данным выборки.

В частности, если предполагаемое распределение — нормальное, то оценивают два параметра (математическое ожидание и среднее квадратическое отклонение), поэтому $r = 2$ и число степеней свободы $k = s - r - 1 = s - 3$.

Так как односторонний критерий более жестко отвергает нулевую гипотезу, чем двусторонний, построим правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α :

$$P\left[\chi^2 > \chi^2_{кр}(\alpha; k)\right] = \alpha.$$

Обозначим значение критерия, вычисленное по данным наблюдений, через $\chi^2_{НАБЛ}$ и сформулируем правило проверки нулевой гипотезы.

Правило: для того чтобы при заданном уровне значимости проверить нулевую гипотезу H_0 (генеральная совокупность распределена нормально), надо сначала вычислить теоретические частоты, а затем наблюдаемое значение критерия:

$$\chi^2_{НАБЛ} = \sum \frac{(n_i - n_i')^2}{n_i}$$

и по таблице критических точек распределения χ^2, по заданному уровню значимости α и числу степеней свободы $k = s - 3$ найти критическую точку $\chi^2_{КР}(\alpha, k)$. Если $\chi^2_{НАБЛ} < \chi^2_{КР}$ — нет оснований отвергнуть нулевую гипотезу, если $\chi^2_{НАБЛ} > \chi^2_{КР}$ — нулевую гипотезу отвергают.